Acta Universitatis Sapientiae

Electrical and Mechanical Engineering

Volume 13, 2021

Contents

Process Optimization of PVDF Piezoelectric Nanofiber Production via Electrospinning

Attila Levente GERGELY[1], József KÁNTOR[2]

Sapientia Hungarian University of Transylvania, Cluj-Napoca,
Faculty of Technical and Human Sciences, Târgu Mureș,
Department of Mechanical Engineering,
e-mail: [1]agergely@ms.sapientia.ro, [2]kantorjozsef@ms.sapientia.ro

Abstract: The aim of this work was to investigate the effect of processing parameters of the electrospinning method on the resulting poly (vinylidene fluoride) (PVDF) fiber diameter. A three factorial Box-Behnken experimental design was employed to study the influence of applied voltage, the capillary-to-collector distance and the applied flow rate on the resulting fiber diameter. We successfully prepared bead-free PVDF nanofibers with fiber diameters ranging from 510-1300 nm. The experimental design analysis did not show significant influence of the studied process parameters under the used boundary conditions on the fiber diameter, thus indicating the robustness of the process.

Keywords: PVDF, electrospinning, nanofiber, piezoelectricity, polymer processing.

1. Introduction

There is a constant need to develop better and more efficient processes for sensor applications. In this respect the use of nanofibers offers high potential in several applications due to their unique properties. The deposition of nanofibers provides a fiber mat, a structure that has a high surface-area-to-volume ratio. The surface-to-volume ratio is inversely proportional to the characteristic cross-sectional dimension of the fibers - usually referred to as fiber diameter, as most fibers have near-circular cross-section. Taking advantage of this property, research groups experimented with fiber mats in multiple domains, some noteworthy applications being filters, scaffolds for cell growth, carriers for drug delivery, catalysis, sensors and actuators [1], [2].

Electrospinning is a widely used method to create polymeric fibers, due to the simplicity of its setup for small scale production under laboratory

conditions. The process involves the placement of a polymer solution droplet into an electric field, while the droplet is in contact with the positive electrode of the instrument. When there is adequate positive charge buildup on the surface of the droplet, it starts to deform, creating the Taylor cone. Eventually the electrostatic forces overcome the surface tension and an electrified jet is created [3]. If new solution is fed into the droplet with the appropriate volumetric flow rate, the Taylor cone and the jet can be maintained at a steady state. The electrified jet travelling from the positive electrode to the grounded electrode (collector) gets elongated by the Coulomb forces and in the process it loses mass by solvent evaporation. At some point the thinning jet becomes unstable due to the increasing volumetric charge density, arranging itself into a coil-like shape and starts a whipping motion. This is called the bending instability, and it causes the jet to travel a much longer distance until reaching the collector than a straight line between the positive electrode and the collector [4]. The elongation of the jet proceeds until it either reaches the collector or dries out enough that the viscoelastic forces become stronger than the Coulomb forces.

The size and morphology of the polymeric fibers produced with electrospinning have a great variety, depending on the materials involved and the process parameters. The fiber sizes range from several micrometers to tens of nanometers, but the most common is the several hundreds of nanometers. The fibers are usually deposited in a random manner (non-woven fiber mats), and extra effort is required to orient the fibers in certain ways. The different fiber mat morphologies can provide different uses [5], [6]. It was found that in keratinocyte cell cultures cell proliferation in a PVA fiber mat with fiber size of 70 nm wasn't different from the control (natural extracellular matrix), while 300 nm fibers resulted in a 100% increase in cell proliferation during a 5-day period. Further increasing the fiber size to 1000 nm caused a decrease in cell proliferation rate compared to 300 nm, but it was still 30% higher than the control [5]. Bian *et al.* published results on filters made of nylon fibers which showed that the removal efficacy of particles (size smaller than 2.5 µm) was close to 98% at ~50 nm fiber size, and gradually dropped to 60% as fiber size increased to 250 nm. [6]

Thus, it is desirable to control the morphological aspects (e.g. fiber size, porosity, orientation, packing density, presence of beads) of the created fiber mats, and to determine which process parameter affects these the most.

Focusing on the fiber size, according to theory, the higher the volumetric charge density in a certain jet segment, the stronger the tensile force, thus the more intensive the elongation. Increasing the voltage increases the volumetric charge density in the solution [7], and in the model system of polyethylene oxide (PEO)/water it was proven that electrospinning at higher charge densities yields thinner fibers [8]. The distance between the positive and the grounded

electrode (needle-collector, NC distance) has multiple ways of interacting with the fiber size. At a given voltage the longer NC distance yields lower electric field strength, and thus causes the weakening of the electrostatic forces [7]. However, a longer NC distance also allows more time for the jets to elongate and for the solvent to evaporate. The rate of solvent evaporation is also crucial as it influences the viscoelastic properties of the jet [3], [9]. Other important factors include solution viscosity, ambient temperature and relative humidity, surface tension, relaxation delay [10].

Published experimental data on the correlation of electrospinning parameters and fiber size is somewhat controversial. For example some studies showed that increasing the applied voltage made the produced fibers thinner [11]–[13], but there are also studies that showed the opposite effect [14]. This indicates that, as of now, there is no generally applicable model for all polymer/solvent systems, and each system has to be studied individually in order to find the optimal electrospinning conditions.

Polyvinylidene fluoride (PVDF) is a semi-crystalline polymer that could exhibit piezoelectric properties [15]. There are five known crystal phases (α, β, γ, δ, ε), [16] and the β-phase and to a lesser extent the γ- and δ-phases contribute to the piezoelectic behavior. The β-phase is the parallel packing of the all-anti configuration of PVDF molecules (also referred to as all-trans in the relevant literature), which is polar in nature and capable of inducing the highest dipole moments out of all the crystal phases [17]. Due to its thermodynamic stability, the α-phase is usually overrepresented in the produced PVDF films. The transition from α- to β-phase was initially achieved with energy-intensive procedures requiring pressures up to 5000 atm and 285°C [17]. More recently electrical poling has been used, that involves the stretching of the sheet at elevated temperature and placing it in an electric field of 10^7 V/m order of magnitude [18].

During electrospinning PVDF undergoes a process similar to poling. The stretching of the fibers and the strong electric field facilitates the formation of β-phase [19]. Experimental data showed that in electrospun fiber mats the β-phase content can reach 85-90% [19], [20]. The electrospinning of PVDF has been successfully used on multiple occasions to create fiber mats that exhibited piezoelectric properties and subsequently devices that incorporated these fiber mats. [21]–[23]

The aim of this work was to determine the significance of the electrospinning process parameters on the resulting PVDF fiber diameter. Thus, a three factorial Box-Behnken experimental design was used, in which the applied voltage, the needle-to-collector distance and the flow rate were selected as factors. In addition, the presence of $β$ and $γ$ crystalline phases in the PVDF

nanofibers were confirmed by FTIR. A simple setup was used to determine the sensitivity of the produced piezoelectric PVDF fiber mats.

2. Experimental

2.1. Materials

Polyvinylidene fluoride (PVDF) (MFI 8g/10min, at 2.16 kg and 230 °C, ASTM D1238, Songhan Plastic Technology), acetone (ACS grade, Sigma Aldrich) and dimethyl formamide (DMF, ACS grade, Sigma Aldrich) were used as received.

2.2. Polymer Solution

20 wt% PVDF in 1:1 acetone:DMF (v:v) solution concentration was selected based on preliminary experiment. To dissolve the PVDF in the solvent mixture ultrasonication was employed at 70 °C for 3 hours, resulting in a clear viscous polymer solution.

2.3. Electrospinning

Electrospinning was carried out on a custom-built apparatus [24]. The PVDF solution was placed into a syringe, and the constant flow rate was ensured by a syringe pump (Fisher Scientific). The syringe was connected to a G21 needle via PTFE tubing. The collector was covered with Al foil and placed at the desired distance from the needle. The positive potential of the DC high voltage supply (Gamma Scientific) was connected to the needle and the collector was grounded. The electrospinning process was carried out at ~21 °C and ~30 % relative humidity.

2.4. Scanning Electron Microscopy (SEM)

The fiber mat morphology was examined with a JEOL JSM-5200 scanning electron microscope at 10 kV potential on neat, not sputter coated, samples. To calculate the fiber diameter at least 50 measurements per sample were averaged. ImageJ software was used to measure the fiber diameters.

2.5. Fourier-Transform Infrared Spectroscopy (FTIR)

Infrared spectra of the PVDF nanofibers were collected using a Frontier, PerkinElmer instrument in ATR (attenuated total reflection) mode between 600 - 1500 cm^{-1} wave number, with a resolution of 1 cm^{-1}.

2.6. Experimental Design

The Box-Behnken experimental design (BBD) was generated with the MiniTab software. The applied voltage (V), the needle-to-collector distance (ND) and the flow rate (F) of the polymer solution were selected as factors. *Table 1* shows the boundary limits determined by scouting experiments to produce a stable electrospinning process. The reason we chose the BBD is that it is almost as powerful as a full factorial design but being a fractional factorial design, it requires significantly fewer experiments. With 3 factors it is 27 runs against 15 in favor of the BBD, and thus more efficient in terms of time commitment and resources. A trade-off is that the BBD is less robust, as each data point is critical to the model.

Table 1. Box-Behnken experimental design limits

Electrospinning parameter	Low (-1)	Mid (0)	High (+1)
Flow rate [mL/h]	0.4	0.7	1
Voltage [kV]	14	17	20
Needle-to-collector distance [cm]	10	15	20

3. Results and Discussion

3.1. Fiber morphology

The fiber morphology and fiber diameter were determined by SEM investigation. The SEM images of some samples are shown in *Fig. 1*. The images as well as the tabulated data in *Table 2* illustrate that out of the 15 experimental condition sets, only three produced PVDF fiber mats without beads. As it is illustrated, beads-on-fiber formation is prominent in the samples. It has been reported that in model systems with PEO low viscosity and high surface tension of the solution, and low charge density (which correlates with low applied voltage) favor the formation of beaded fibers [8]. Since there weren't any viscosity or surface tension measurements performed on the PVDF solutions, it can only be proposed that the same applies here as well. Based on our observations the viscosity of the used PVDF solutions was high enough for continuous fiber formation (as the SEM images show). The produced fibers were smooth surfaced and have circular cross section.

PVDF2

PVDF3

PVDF4

PVDF5

PVDF9

PVDF11
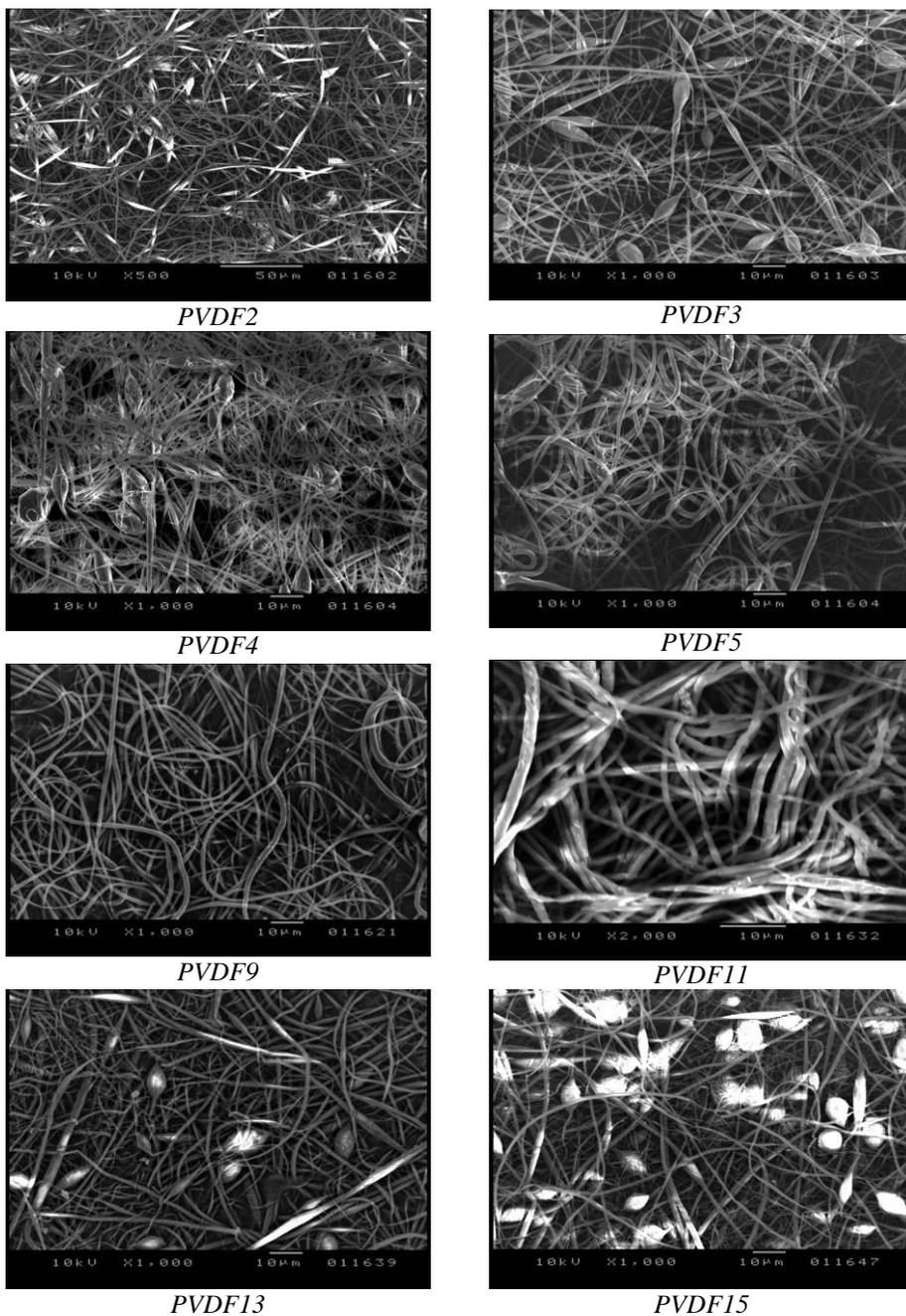
PVDF13

PVDF15

Figure 1: Representative SEM images of the tested samples

Table 2 presents the generated parameter sets for the BBD. Samples named PVDF5, PVDF9 and PVDF11 are the ones that did not contain beads. It can also be noted that the fiber diameters range from 510 nm to 1300 nm

Table: Fiber diameter and standard deviations.

Name	Needle-collector distance [cm]	Volumetric flow rate [mL/h]	Voltage [kV]	Fiber size, d [μm]	Stdev. d [μm]	Bead
PVDF1	15	0.7	17	0.70	0.36	Yes
PVDF2	15	0.4	14	1.30	0.55	Yes
PVDF3	20	0.7	14	0.93	0.52	Yes
PVDF4	15	1	14	0.73	0.45	Yes
PVDF5	10	1	17	1.09	0.44	No
PVDF6	15	0.7	17	0.92	0.38	Yes
PVDF7	15	0.7	17	0.80	0.36	Yes
PVDF8	10	0.7	14	0.51	0.25	Yes
PVDF9	10	0.7	20	0.92	0.23	No
PVDF10	15	1	20	1.07	0.36	Yes
PVDF11	20	0.7	20	1.24	0.36	No
PVDF12	20	0.4	17	1.11	0.44	Yes
PVDF13	15	0.4	20	1.06	0.46	Yes
PVDF14	10	0.4	17	0.92	0.25	Yes
PVDF15	20	1	17	0.76	0.40	Yes

3.2. Experimental design analysis

To investigate the effect of the electrospinning parameters on the fiber size and spread a 3-factor BBD was created. *Tables 3* and *4*, show the respective ANOVA tables.

Table 3: ANOVA table for fiber diameter of the studied fiber mats

Source	DF	Adj. SS	Adj MS	F-value	p-value
Linear	**3**	**0.13734**	**0.04578**	**1.22**	**0.393**
dist	1*	0.01051	0.01051	0.28	0.619
flow	1*	0.0931	0.0931	2.49	0.176
volt	1*	0.030637	0.030637	0.82	0.407
Square	3	0.103227	0.034409	0.92	0.495
Interaction	3	0.154200	0.051400	1.37	0.352
Residual Error	5	0.187167	0.037433		
Total	14				
R-Sq = 70.85% R-Sq(adj) = 18.38% *-accounted for in the DF of Linear source					

The data in *Table 3* shows, that for the fiber diameter, none of the p-values were lower than 0.05 (the lowest being 0.176) and as such the effects of the selected electrospinning parameters indicated no statistical significance in the 95% confidence interval. A study done by Motamedi *et al.* showed that the diameter of PVDF nanofibers electrospun from dimethylacetamide/acetone mixture increased with increasing voltage (300 nm at 10 kV to 855 nm at 20 kV) and decreased with increasing distance (644 nm at 14 cm to 264 nm at 20 cm) [25]. On the other hand, according to the results published by Ribeiro *et al.* average fiber sizes spun from DMF solution decreased from 495 to 403 nm when they increased the voltage from 15 to 30 kV [20]. Neither publication provides statistical data, so the statistical significance of the results is unknown. While not observed for all material/solvent systems (as the first example shows, [25]), it is expected in general that a higher voltage through an increased volumetric charge density and larger tensile forces would yield thinner fibers [8]. As the opposite effect, larger distances would yield thicker fibers because of the weaker electric field and lower charge density [7]. It is theorized that in the case when fibers get thicker with increasing voltage/ higher charge density it is because more solution is drawn from the source and the increased volumetric flow rate results in a thicker jet [14]. In the previously mentioned study with PVDF there wasn't a clear correlation between volumetric flow rate and deposited fiber size [25].

Table 4: ANOVA table for fiber diameter standard deviation of the studied fiber mats

Source	DF	Adj. SS	Adj MS	F-value	p-value
Linear	**3**	0.031937	0.010646	**2.28**	**0.197**
dist	1*	0.025345	0.025345	5.42	0.067
flow	1*	0.001186	0.001186	0.25	0.636
volt	1*	0.001035	0.001035	0.22	0.658
Squared	3	0.028652	0.009551	2.04	0.227
Interaction	3	0.018125	0.006042	1.29	0.373
Residual Error	5	0.023392	0.004678		
Total	14				
R-Sq = 81.21% R-Sq(adj) = 47.39% *-accounted for in the DF of Linear source					

For the standard deviation of the fiber size, *Table 4*, the distance had a p-value 0.067, which is close to 0.05. While technically not significant by to the chosen convention of the <5% probability for the null hypothesis, it still might be an indication that the widening of the fiber size distribution with the increasing needle-collector distance is not entirely due to randomness. This trend can be observed in the data of Motamedi *et al.* as well, even though the deviation in their samples was roughly an order of magnitude lower [25]. A possible explanation is found in the relationship between solvent evaporation

and the needle-collector distance. It is known that the elongation of the fibers is limited by the drying process [6]. In each electrospinning setup, especially in the more unstable systems there is a variation in the initial jet diameter (i.e. close to the Taylor cone). The initially thin jet segments dry out quickly, and after that point no additional distance will affect the fiber size. On the other hand, for the initially thick jet segments there is more time for the solvent to evaporate if the distance is longer, and thus the result will be a narrower size distribution.
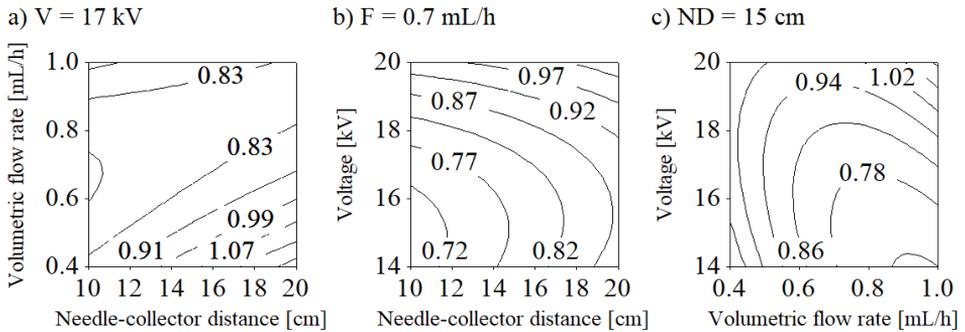
Figure 2: Contour plots of the mean fiber diameter, d [μm], versus a) F-ND, b) V-ND, and c) V-F when the respective third parameter was held at its middle value

While the statistical analysis indicated no significance for the effect of the electrospinning parameters on fiber diameter, some conclusions regarding the process can be drawn from the observations. The contour plots seen in *Fig. 2* show that in the studied region increasing the voltage caused an increase in average fiber size from 600 to 900 nm at F = 0.7 mL/h and ND = 10 cm (*Fig. 2b*), and from 700 to 1000 nm at F = 1 mL/h and ND = 15 cm (*Fig. 2c*). The increase in distance also resulted in thicker fibers in general (*Fig. 2a-b*). Increasing the flow rate caused the average fiber size to decrease. At V = 14 kV and ND = 15 cm it went down from 1100 nm to 700 nm (*Fig. 2c*) and at V = 17 kV, ND = 20 cm from 1200 nm to 800 nm (*Fig. 2a*), however in both cases it seems to have reached a local minimum, after which the fiber size would have started increasing again if the flow rates were to become higher. The low distance favored more uniform fibers, and at higher voltages the bead formation became less prevalent. The most stable condition was at 10 cm distance, 20 kV, and 0.7 mL/h (sample PVDF9), with the fiber mat having submicron fiber size, one of the narrowest fiber size distributions, and no beads.

3.3. Piezoelectric properties

The presence of β and γ crystalline phases, thus the piezoelectric properties of the samples were confirmed by FTIR spectroscopy. A representative FTIR spectrum can be seen in *Fig. 3*. Based on the work of Cai *et al.* the β and γ crystalline phase content relative to the α phase was calculated for each sample, using eq. (1):

$$F_{EA} = \frac{I_{EA}}{\left(\frac{K_{840}}{K_{763}}\right) I_{763} + I_{EA}} \tag{1}$$

where F_{EA} represents the β and γ crystalline phase percentage, I_{EA} and I_{763} represents the absorbance values at 840 and 763 cm^{-1} respectively, whereas K_{840} = 7.7*10^4 and K_{763} = 6.1*10^4 cm^2/mol are the absorbance coefficients at the particular wave number [26]. The FTIR spectra was normalized at ware number 1072 cm^{-1}, since the absorbance of this band is influenced only by the thickness of the sample and not the crystalline structure [27].
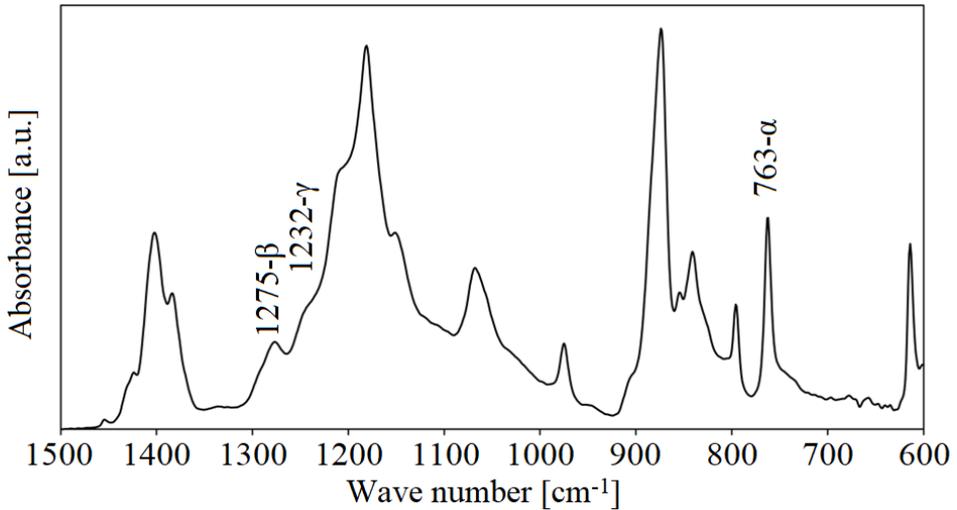
Figure 3: Representative FTIR spectrum

The calculated β and γ crystalline phases for all the samples were within 36-42% with respect to the α phase. To test the piezoelectric properties of the samples a test cell was built in which the samples were placed between two masked copper plates and were subjected to compression forces, while the generated voltage was measured. The samples generated potential difference in response to the loading, however the highest sensitivity (generated

voltage/applied force) of 0.027 mV/N was obtained for sample PVDF11 with 11 μm thickness. The control, 100 μm thick PVDF film produced 0.063 mV/N sensitivity. Thus, the sensitivity of the prepared PVDF nanofiber samples was in the same order of magnitude as the sensitivity of the commercially available PVDF film.

4. Conclusion

In summary, we successfully prepared bead-free smooth surfaced PVDF fiber mats with piezoelectric properties. The 3-factorial Box-Behnken experimental design did not show significant influence of the studied electrospinning parameters on the resulting fiber diameter, thus indicating the robustness of the process. FTIR study confirmed the presence of the electroactive β and γ crystalline phases in all samples in the range of 36-42 % with respect to the α phase. The sensitivity of the electrospun piezoelectric PVDF nanofiber samples was comparable to the control, a commercially available PVDF film.

Acknowledgements

References

[1]   Xue, J., Wu, T., Dai, Y., and Xia, Y., "Electrospinning and electrospun nanofibers: Methods, materials, and applications", *Chem. Rev.*, vol. 119, no. 8, pp. 5298–5415, 2019.
[2]   Sipos, E. *et al.*, "Preparation and characterization of fenofibrateloaded PVP electrospun microfibrous sheets", *Pharmaceutics*, vol. 12, no. 7, 2020.
[3]   Yarin, A. L., Koombhongse, S., and Reneker, D. H., "Taylor cone and jetting from liquid droplets in electrospinning of nanofibers", *J. Appl. Phys.*, vol. 90, no. 9, pp. 4836–4846, 2001.
[4]   Yarin, A. L., Koombhongse, S., and Reneker, D. H., "Bending instability in electrospinning of nanofibers", *J. Appl. Phys.*, vol. 89, no. 5, pp. 3018–3026, 2001.
[5]   Pelipenko, J., Kocbek, P., and Kristl, J., "Nanofiber diameter as a critical parameter affecting skin cell response", *Eur. J. Pharm. Sci.*, vol. 66, pp. 29–35, 2015.
[6]   Bian, Y., Wang, S., Zhang, L., and Chen, C., "Influence of fiber diameter, filter thickness, and packing density on PM2. 5 removal efficiency of electrospun nanofiber air filters for indoor applications", *Build. Environ.*, vol. 170, p. 106628, 2020.

[7]   Theron, S. A., Zussman, E., and Yarin, A. L., "Experimental investigation of the governing parameters in the electrospinning of polymer solutions", *Polymer (Guildf).*, vol. 45, no. 6, pp. 2017–2030, 2004.

[8]   Fong, H., Chun, I., and Reneker, D. H., "Beaded nanofibers formed during electrospinning", *Polymer (Guildf).*, vol. 40, no. 16, pp. 4585–4592, 1999.

[9]   Tripatanasuwan, S., Zhong, Z., and Reneker, D. H., "Effect of evaporation and solidification of the charged jet in electrospinning of poly (ethylene oxide) aqueous solution", *Polymer (Guildf).*, vol. 48, no. 19, pp. 5742–5746, 2007.

[10] Thompson, C. J., Chase, G. G., Yarin, A. L., and Reneker, D. H., "Effects of parameters on nanofiber diameter determined from electrospinning model", *Polymer (Guildf).*, vol. 48, no. 23, pp. 6913–6922, 2007.

[11] Amiraliyan, N., Nouri, M., and Kish, M. H., "Effects of some electrospinning parameters on morphology of natural silk-based nanofibers", *J. Appl. Polym. Sci.*, vol. 113, no. 1, pp. 226–234, 2009.

[12] Beachley, V., and Wen, X., "Effect of electrospinning parameters on the nanofiber diameter and length", *Mater. Sci. Eng. C*, vol. 29, no. 3, pp. 663–668, 2009.

[13] Nasouri, K., Bahrambeygi, H., Rabbi, A., Shoushtari, A. M., and Kaflou, A., "Modeling and optimization of electrospun PAN nanofiber diameter using response surface methodology and artificial neural networks", *J. Appl. Polym. Sci.*, vol. 126, no. 1, pp. 127–135, 2012.

[14] Jia, Z., Li, Q., Liu, J., Yang, Y., Wang, L., and Guan, Z., "Preparation and properties of poly (vinyl alcohol) nanofibers by electrospinning", *J. Polym. Eng.*, vol. 28, no. 1–2, pp. 87–100, 2008.

[15] Kawai, H., "The piezoelectricity of poly (vinylidene fluoride)", *Jpn. J. Appl. Phys.*, vol. 8, no. 7, p. 975, 1969.

[16] Lovinger, A. J., "Annealing of poly (vinylidene fluoride) and formation of a fifth phase", *Macromolecules*, vol. 15, no. 1, pp. 40–44, 1982.

[17] Hasegawa, R., Kobayashi, M., and Tadokoro, H., "Molecular conformation and packing of poly (vinylidene fluoride). Stability of three crystalline forms and the effect of high pressure", *Polym. J.*, vol. 3, no. 5, pp. 591–599, 1972.

[18] Ramos, M. M. D., Correia, H. M. G., and Lanceros-Mendez, S., "Atomistic modelling of processes involved in poling of PVDF", *Comput. Mater. Sci.*, vol. 33, no. 1–3, pp. 230–236, 2005.

[19] Gee, S., Johnson, B., and Smith, A. L., "Optimizing electrospinning parameters for piezoelectric PVDF nanofiber membranes", *J. Memb. Sci.*, vol. 563, pp. 804–812, 2018.

[20] Ribeiro, C., Sencadas, V., Ribelles, J. L. G., and Lanceros-Méndez, S., "Influence of processing conditions on polymorphism and nanofiber morphology of electroactive poly (vinylidene fluoride) electrospun membranes", *Soft Mater.*, vol. 8, no. 3, pp. 274–287, 2010.

[21] Lang, C., Fang, J., Shao, H., Ding, X., and Lin, T., "High-sensitivity acoustic sensors from nanofibre webs", *Nat. Commun.*, vol. 7, no. 1, pp. 1–7, 2016.

[22] Wang, Y. R., Zheng, J. M., Ren, G. Y., Zhang, P. H., and Xu, C., "A flexible piezoelectric force sensor based on PVDF fabrics", *Smart Mater. Struct.*, vol. 20, no. 4, p. 45009, 2011.

[23] Fang, J., Wang, X., and Lin, T., "Electrical power generator from randomly oriented electrospun poly (vinylidene fluoride) nanofibre membranes", *J. Mater. Chem.*, vol. 21, no. 30, pp. 11088–11091, 2011.

[24] Gergely, A., Kántor, J., Bitay, E., and Biró, D., "Electrospinning of Polymer Fibres Using Recycled PET", *Acta Mater. Transylvanica*, vol. 2, no. 1, pp. 19–26, 2019.

[25] Motamedi, A. S., Mirzadeh, H., Hajiesmaeilbaigi, F., Bagheri-Khoulenjani, S., and Shokrgozar, M., "Effect of electrospinning parameters on morphological properties of PVDF nanofibrous scaffolds", *Prog. Biomater.*, vol. 6, no. 3, pp. 113–123, 2017.

[26] Cai, X., Lei, T., Sun, D., and Lin, L., "A critical analysis of the α, β and γ phases in poly (vinylidene fluoride) using FTIR", *RSC Adv.*, vol. 7, no. 25, pp. 15382–15389, 2017.

[27] Benz, M. and Euler, W. B., "Determination of the crystalline phases of poly (vinylidene fluoride) under different preparation conditions using differential scanning calorimetry and infrared spectroscopy", *J. Appl. Polym. Sci.*, vol. 89, no. 4, pp. 1093–1100, 2003.

Investigation of the Thermodynamic Characteristics of the Ester Oil and R152a, R125, R134a and R123 Refrigerant Mixtures

Robert Santa

University of Dunaújváros,
Department of Mechanical Engineering,
e-mail: santar@uniduna.hu

Abstract: The presence of the lubricant POE 68 in the refrigeration system changes the thermophysical properties of the refrigerant, which significantly affects the heat transfer and the hydraulic processes. The purpose of this research is to investigate the thermodynamic properties of the R152a, R125, R134a and R123 refrigerant and POE 68 lubricant mixtures at different temperatures from 233,15 to 313,15 K. There have been investigated the values of the densities and kinematic viscosities of the mixtures at different concentrations (100%, 90%, 80%). Finally, it was found that, the density and kinematic viscosity of the R123/POE 68 mixture were most affected by the change of the concentration.

Keywords: Heat pump, refrigerant, oil, mixture, density, viscosity

1. Introduction

The modelling of the heat pumps is based on decades of experience as both of stationary [1], [2] and the non-stationary [3], [4] mathematical models. The mathematical models are fraught with a numerous omissions and uncertainties.

The accuracy of the heat pumps models is greatly affected by the accuracy of the pressure drop correlations which are used to determine the pressure drop of the refrigerant flowing in the tubes and the heat transfer correlations [5], [6] which are used to determine the heat transfer coefficients. In the literature, there are a large number of correlations of heat transfer and pressure drop, whose accuracy, validity is uncertain. Last, but not least, a considerable number of them ignore the effect of the oil in the refrigerant, although refrigerant oil is an essential requirement for the operation of the cycle. This study investigated the changes of the density and kinematic viscosity values of the refrigerant R152a,

R125, R134a, R123 and the POE68 polyester oil mixtures as a function of the temperatures. The analysis was performed at high refrigerant concentrations in the mixture. The concentrations were 100%, 90% and 80 %, where 100 % represents the pure refrigerant.

2. Mathematical model

Due to the complexity of the Navier-Stokes differential equations [7] used to describe the flowing refrigerant, it is extremely difficult to obtain an analytical solution. The form of the differential equation used here is:

$$\frac{\mathrm{d}u}{\mathrm{d}t} + u \cdot \nabla u = -\nabla p + \rho \cdot g + \mu \cdot \nabla^2 u \tag{1}$$

where the significance of the terms is the following:

- ∇p pressure gradient (fluid flows in the direction of largest change in pressure);
- $\rho \cdot g$ body force term (external forces, that act on the fluid);
- $\mu \cdot \nabla^2 u$ diffusion term (for a Newtonian fluid, viscosity operates as a diffusion of momentum).

However, in the equations, the resulting characteristic members can be related to each other as ratios, resulting ratios will be dimensionless expressions. The numerical values of the dimensionless expressions – similarity criteria – thus obtained are independent of the unit system used. In the formulation of hydrodynamic similarity criteria, the force of inertia is chosen as the reference basis.

The characteristic terms of the equation [1] can be primed, related to each other, using ratios. It is useful to turn the equation [1] into a dimensionless form. In order to formulate the hydrodynamic similarity criteria, the force of inertia is used as a reference value.

If all members are divided by this value, the result is the dimensionless criteria. The dimensional analysis method is also used to determine the amount of heat transferred during convective heat flow and the heat transfer coefficient.

In general, the heat transfer coefficient of the refrigerant can be determined as a function of the Reynolds number (Re), Prandtl number (Pr) and Grashof number (Gr):

$$\alpha = \frac{\lambda}{d}, f(\mathrm{Re}, \mathrm{Pr}, \mathrm{Gr}) \tag{2}$$

From the above equations it is clear that, in addition to the flow parameters, the characteristics of the refrigerant, i.e. the density and viscosity of the

refrigerant also play a decisive role. In order to analyze the heat pump cycle, it is necessary to determine the thermodynamic parameters of the function lubricant POE68 in order to be able to define thermal and hydraulic processes.

The mathematical equations (3 - 10) are used to compute the kinematic viscosity and the density of the mixtures of 80%, 90% concentration solutions and of the pure 100% refrigerant. The equations are suitable for high refrigerant concentrations and temperatures between -40 °C and 40 °C.

Algebraic equations are used for determining the density of the R134a/POE 68 polyester (pentaerythritol) Ester Mixed Acid and the R123/POE 68 pentaerythritol Ester Mixed Acid [8]:

$$\rho = \left(a_1 + a_2 \cdot T_r + a_3 \cdot T_r^2\right) + \omega \cdot \left(a_4 + a_2 \cdot T_r + a_6 \cdot T_r^2\right) + \omega^2 \cdot \left(a_7 + a_8 \cdot T_r + a_9 \cdot T_r^2\right)$$

(3)

where,

$$T_r = 1 - \frac{T}{T_C}$$

Kinematic viscosity is given by

$$\log \vartheta = \left(a_1 + \frac{a_2}{T} + \frac{a_3}{T^2}\right) + \omega \cdot \left(a_4 + \frac{a_5}{T} + \frac{a_6}{T^2}\right) + \omega^2 \cdot \left(a_7 + \frac{a_8}{T} + \frac{a_9}{T^2}\right)$$

(4)

and ω is the refrigerant mass fraction.

The following algebraic equations are used for the determination of density and kinematic viscosity of the R125/POE 68 and the R152a/POE 68 pentaerythritol Ester Mixed Acid [8]:

$$\rho_{100} = a_1 + a_2 \cdot T + a_3 \cdot T^2$$

(5)

$$\rho_{90} = a_4 + a_5 \cdot T + a_6 \cdot T^2$$

(6)

$$\rho_{80} = a_7 + a_8 \cdot T + a_9 \cdot T^2$$

(7)

$$\log \vartheta_{100} = a_1 + \frac{a_2}{T} + \frac{a_3}{T^2}$$

(8)

$$\log \vartheta_{90} = a_4 + \frac{a_5}{T} + \frac{a_6}{T^2}$$

(9)

$$\log \vartheta_{80} = a_7 + \frac{a_8}{T} + \frac{a_9}{T^2}$$

(10)

The coefficients are found in [8] and the subscript 100, 90, 80 refer to the mass fraction of the refrigerant.

3. Results and discussion

The figures below summarize the density and kinematic viscosity of the refrigerants R152a, R125, R134a and R123 and of the ester oil lubricant POE 68 as a function of the temperature.

Fig. 1 and *Fig. 2* show the density and kinematic viscosity of a mixture of R152a refrigerant and POE68 oil. In the first figure it can be observed that below the limit of 273.13 K, influence of the concentration on the density, between 80% and 100% is practically insignificant. However, it can also be seen that as the temperature increases, the value of the density of the mixture also decreases considerably.
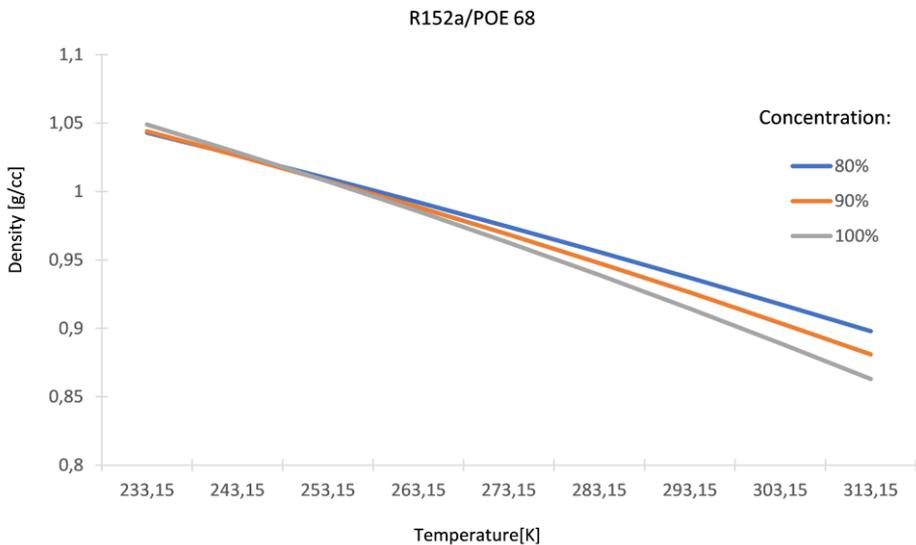
Figure 1: Density of the R152a/POE 68 as a function of the temperature

It can be noticed that at low temperatures 233.15 K and 80% concentration, the kinematic viscosity of the mixture increases significantly, while at higher temperatures 313.5 K, the difference in viscosity values presents less importance at different concentrations.
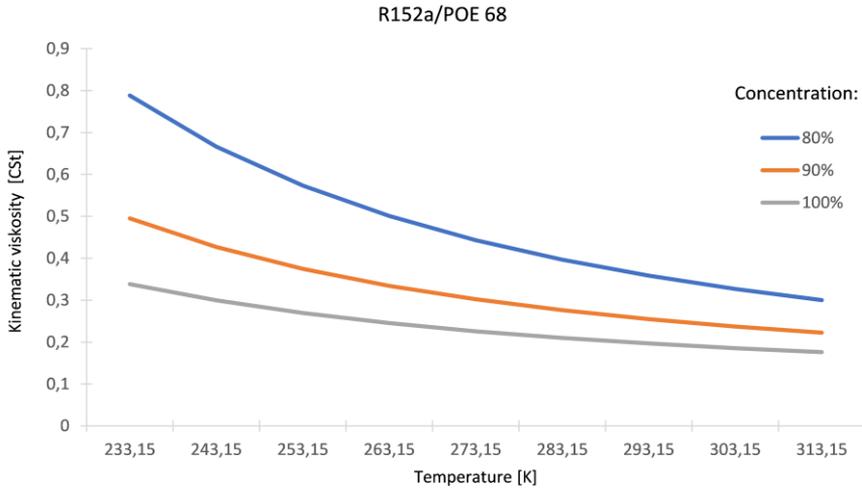
Figure 2: Kinematic viscosity of the R152a/POE 68 as a function of the temperature

The density and the kinematic viscosity of the mixture of R125 and POE68 can be seen in Figures 3 and 4.
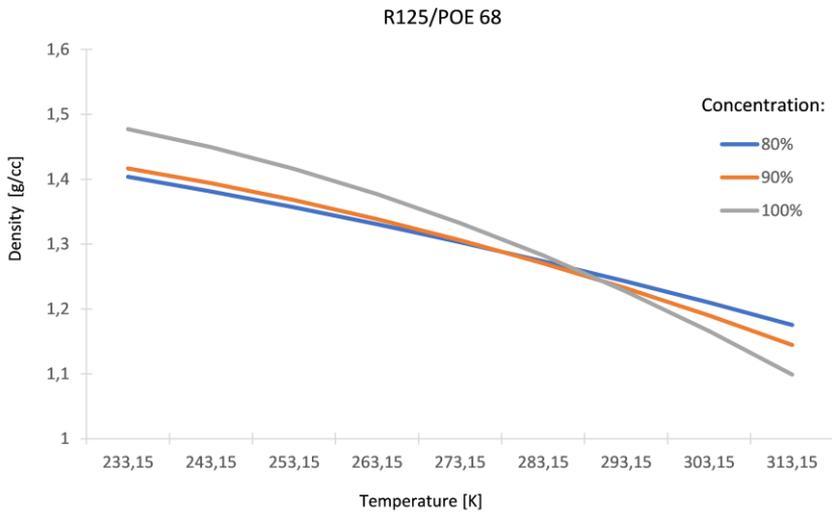
Figure 3: Density of the R125/POE 68 as a function of the temperature

The figure shows that at 283.13 K the density of the mixture is nearly constant for all three concentrations at 80%, 90% and 100%. It can also be seen that as the temperature decreases, and at 100% concentration - pure refrigerant, the value of the refrigerant density is higher than at 80% and 90% concentration, while at a higher temperature of 313.15K the reverse of the phenomenon is observed.

Considering the kinematic viscosity, it is demonstrated that the viscosity value for a pure 100% refrigerant is nearly the same over the temperature range investigated, while for the 80% mixture the viscosity value increases significantly for low temperatures of 233.15 K.
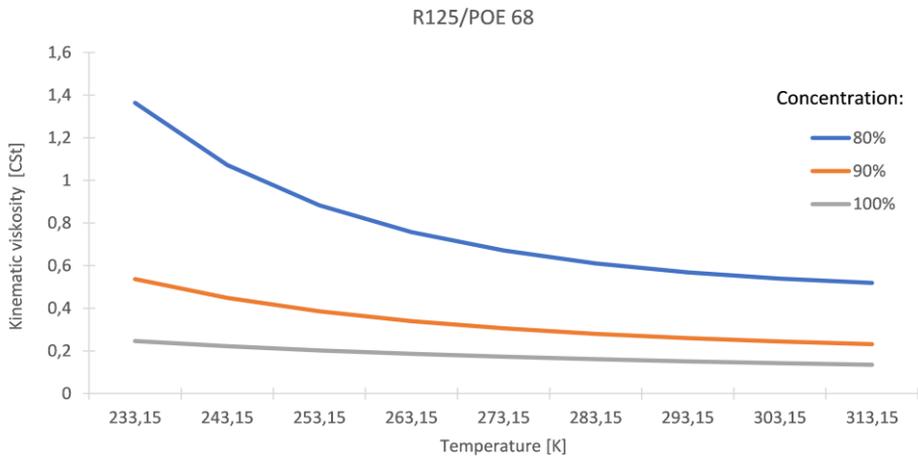
Figure 4: Kinematic viscosity of the R125/POE 68 as a function of the temperature

The density and kinematic viscosity of the mixture of R134a and POE 68 are shown in *Fig. 5* and *Fig. 6*. It can be seen that as the temperature decreases, the density value increases steadily for 80% and 90% concentration, but the highest density values are observed for the pure, i.e., 100% refrigerant.
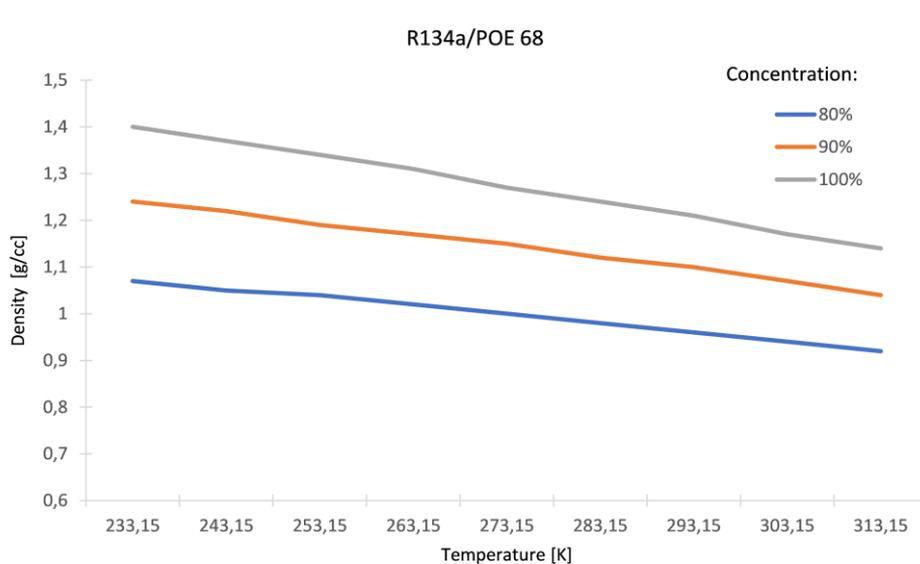
R134a/POE 68

Figure 5: Density of the R134a/POE 68 as a function of the temperature

Fig. 6 shows that at a low temperature of 233.15 K, the kinematic viscosity of a mixture with a concentration of 80% is five times greater than those of the pure refrigerant, whereas at higher temperatures 313.15 K it decreases to nearly the double.
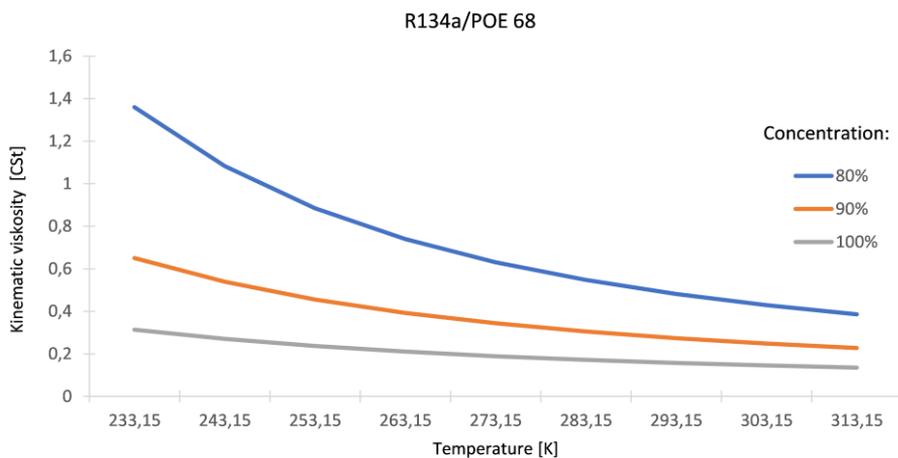
R134a/POE 68

Figure 6: Kinematic viscosity of the R134a/POE 68 as a function of the temperature

The density and kinematic viscosity of the refrigerant R123 and the lubricant POE 68 are summarized in *Fig. 7* and *Fig. 8*. It can be deduced that as the temperature decreases, the density, and the kinematic viscosity values of all three mixtures increase and show the same trend as R134a/POE 68, but presenting larger numerical values, for both thermodynamic parameters.
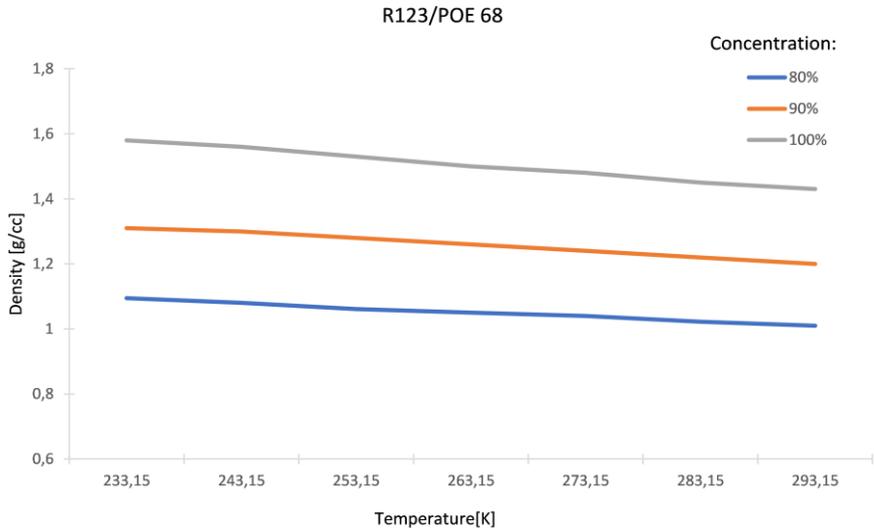
Figure 7: Density of the R123/POE 68 as a function of the temperature
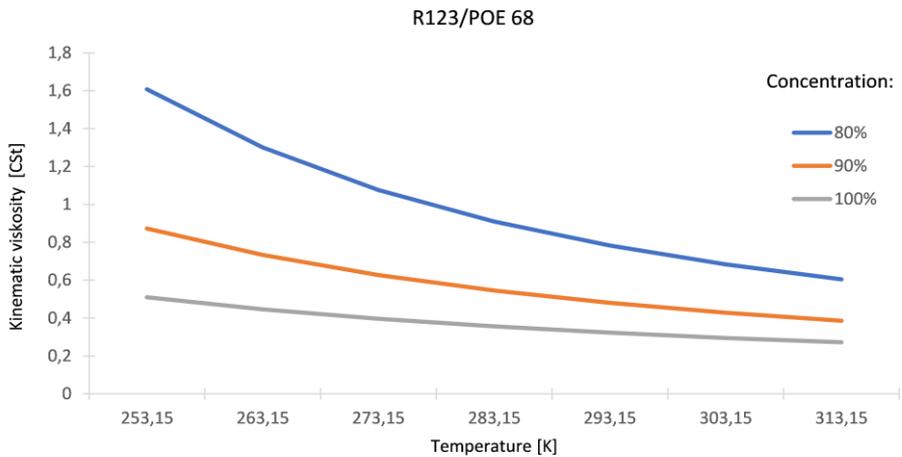
Figure 8: Kinematic viscosity of the R123/POE 68 as a function of the temperature

The following figures show the effect of oil on the change of the heat transfer coefficient by way of a numeric simulation example. In the convective condensation heat transfer correlations [9], [10] the 80% and 90% mixture concentrations were taken into account.

The range of simulation conditions:

- Refrigerant: R134a;

- Vapor quality: 0.1 – 0.9 [-];

- Mass flux: 125, 250 [kg/m^{2}s];

- Concentration: 80% 90%, 100%.
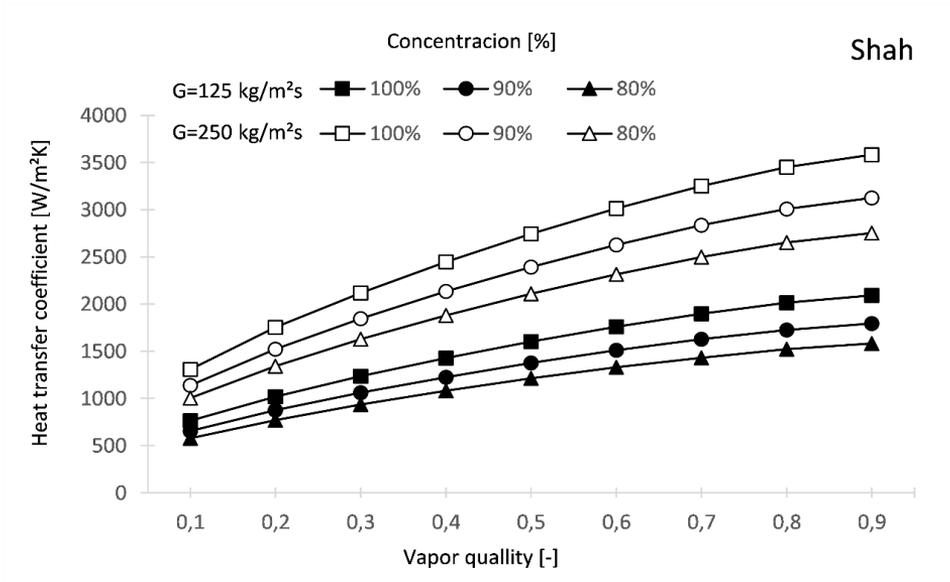
Figure 9: Heat transfer coefficient based on the Shah correlation [9]
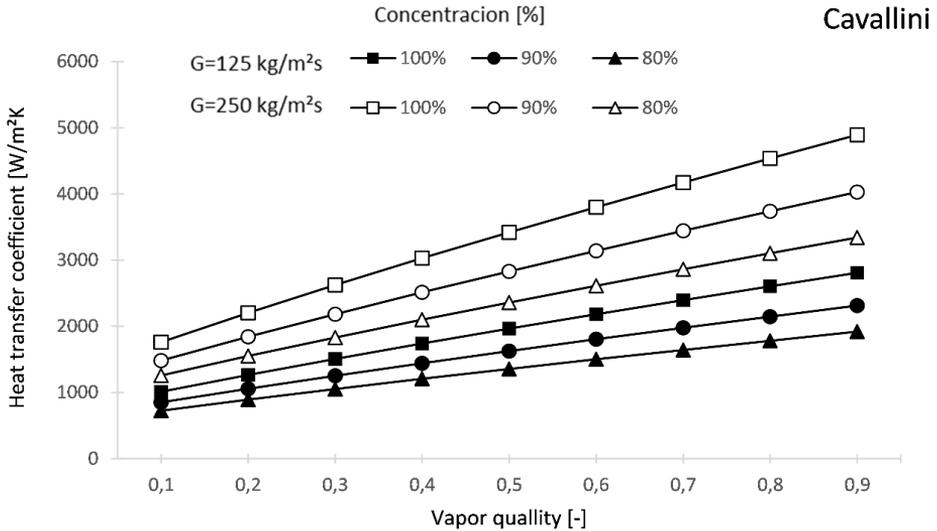
Figure 10: Heat transfer coefficient based on the Cavallini correlation [10]

4. Conclusion

The thermodynamic properties, density, and kinematic viscosity, of the R152a, R125, R134a and R123 refrigerants and POE 68 ester oil mixture, were investigated as a function of refrigerant temperature.

It was found that the density values of the R134a/POE 68 and R123/POE 68 mixtures showed the same trend, with lower density values at 80% concentration than those of the pure refrigerant.

It must be also emphasized that for R152a/POE 68 and R125/POE 68 the density values were almost identical for 80%, 90% and 100% concentration.

For the mixtures R152a/POE 68, R125/POE 68, R134a/POE 68 and R123/POE 68, the change in kinematic viscosity values was the same as a function of temperature. Experiments highlighted that the mixture R123/POE68 had the highest density and kinematic viscosity of all investigated mixtures: 1.6 g/cc respectively 1.5 CSt. Furthermore, it was determined that with the increase of concentration, all four mixtures presented a strong increasing trend of the kinematic viscosity.

Besides this, it must be highlighted that the density values of all mixtures were significantly lower than the density of the pure refrigerant.

In conclusion, increasing the proportion of ester oil in the mixture concentration had a negative effect on the amount of heat transferred during

convective heat flow and also on the heat transfer coefficient Eq. (2). *Fig. 9* and *Fig. 10* showed the effect of the mixture of refrigerant and oil on the values of the heat transfer coefficient as a function of vapor quality and heat mass flux. Two condensation heat transfer models were investigated, namely the Shah and Cavallini correlations. In both cases, it can be seen that as the proportion of oil increases, the values of the heat transfer coefficient decrease.

References

[1]   Santa, R., "Comparative Analysis of Heat Pump System with IHX Using R1234yf and R134a", *Periodica Polytechnica Mech. Eng.,* vol. 65, pp. 1−11, 2021.

[2]   Kassai, M., "Investigation of the thermal behavior and energy consumption of refrigeration systems", *Thermal Science*, vol. 25, no. 1 Part A, pp. 73–83, 2021.

[3]   Sánta, R., Garbai, L., Fürstner, I., "Optimization of heat pump system", *Energy,* vol. 89 pp. 45–54, 2015.

[4]   Sánta, R., Garbai, L., Fürstner, I., "Numerical investigation of the heat pump system", *J. Thermal Cal*., vol. 130, no. 2, pp. 1133–1144, 2017.

[5]   Santa, R., "The Analysis of Two-phase Condensation Heat Transfer Models Based on the Comparison of Boundary Condition", *Acta Poly. Hun*., vol.9, no. 6, pp. 167–180, 2012.

[6]   Santa, R., Garbai, L., "Measurement testing of heat transfer coefficients in the evaporator and condenser of heat pumps", *J. Thermal Cal.*, vol. 119, no. 3, pp. 2099–2106, 2015.

[7]   Garbai, L., "Csőhálózatok hidraulikája: Állandósult áramlás", Akadémiai Kiadó, Budapest, 2018

[8]   Thomas, J., Bruno, T.J., Fortin, M. L., Huber, A.L., Eric W. L., Elisabeth M., Mark O. M., Stephanie L. O., Richard A. P., Kimberly, N. U., Jason A. W., "Thermophysical Properties of Polyol Ester Lubricants", *NIST Director and Undersecretary of Commerce for Standards and Technology*, 2019.

[9]   Shah, M. M., "A General Correlation for Heat Transfer during Film Condensation inside Pipes", *International Journal of Heat and Mass Transfer* 22, no.4, pp: 547–56, 1979.

[10] Cavallini, A., and R. Zecchin, "A dimensionless correlation for heat transfer in forced convection condensation", in *Proceedings of the Fifth International Heat Transfer Conference, Tokyo*, 1974, pp. 309–313.

Kinematic and Dynamic Modeling of the Rotary Harrow

Ferenc TOLVALY-ROSCA[1], Judit PÁSZTOR[2], Zoltán FORGÓ[3]

Sapientia Hungarian University of Transylvania, Cluj-Napoca,
Faculty of Technical and Human Sciences, Târgu Mureș,
Department of Mechanical Engineering,
e-mail: [1]tferi@ms.sapientia.ro, [2]pjudit@ms.sapientia.ro, [3]zforgo@ms.sapientia.ro

Abstract: The actual energy situation and the significant energy demand of agricultural production require the exploration and analysis of the general laws of the main processes in the field of energy, which is part of the borderline between the technical sciences and the agricultural sciences. Tillage machines change the physical properties of the soil, but at the same time the mechanical properties of the soil react to the implement. One of the basic works of seedbed preparation is harrowing. In this work we study the kinematics and dynamics of the active rotary harrow implements, taking into account the cutting resistance of the soil. The results obtained with the developed modeling methods, give reliable approximations to the experienced tilling processes.

Keywords: Rotary harrow, kinematics, dynamics, assembly model.

1. Introduction

Tillage is a mechanical intervention, the aim of which is to create favorable soil conditions for the crops to grow. Tillage consists of basic operations and seedbed preparation. The basic operation is the deepest, the seedbed preparation is shallower, but the most demanding intervention.

The purpose of seedbed preparation is rapid germination, which is facilitated by a soil condition in which the circulation of moisture, air and heat is favorable. During seedbed preparation, the top layer of soil is loosened, crushed, levelled and compacted at the depth of sowing. It is necessary to choose a method and a tool by which the conditions of good germination can be realized cost-effectively with little intervention, without damaging the soil structure. Machines that perform these jobs in one run at the same time are common. These machines are usually equipped with active tillage implements, which are driven by the tractor's PTO shaft. One such work machine is the rotary harrow [1], [2].

The rotary harrow is the seedbed preparation work machine. In the course of its work, it shreds the soil. It also stirs up, flattens and mixes vigorously while shreds. The rotary harrow is power-driven, its implements are active, and they are also driven from the power take-off shaft (PTO), during towing. Thus, the power requirement of the traction work is lower, the slip loss during towing is smaller, therefore it can be used even in wetter ground conditions [1], [3].

The main goal of the machine operation is to provide the required seedbed with the lowest possible number of turns and favorable energy consumption [4], [5].

Using a nonlinear mathematical modeling and simulation of the systems [6] [7], the movement trajectories of a work tool can be studied, [1], [3], [8], [9], [10], [11].

Dynamic models can be used to study the dynamic behaviour [12] and the energy demand [13] of a work tool. In the present paper we perform kinematic and dynamic modelling of the rotary harrow teeth.

2. Materials and methods

We perform kinematic and dynamic modeling of the rotary harrow. To do this, we made a simplified assembly model of the rotary harrow, which is used for dynamic modeling. The trajectories of the harrow teeth are generated using mathematical modeling, the forces acting on one harrow tooth are determined, then they are used to make computer simulations for dynamic and energetic studies.

2.1. The simplified assembly model

The working tool of the rotary harrow is the tine. Since the rotation direction of the adjacent work units is opposite, two types of tines are used according to each rotation directions (*Fig. 1*).

Figure 1: The solid models of the tines (left) and a working unit with two tines (right)

The implements are driven from the power take-off shaft of the power unit, via a deceleration gearbox (*Fig. 2*). The PTO shaft spindle speed is conventional, 540 rpm. The drive chain is designed so that the rotational movement from the tractor-PTO is transmitted by gears to the vertical axes, so that the adjacent implements rotate in opposite directions.
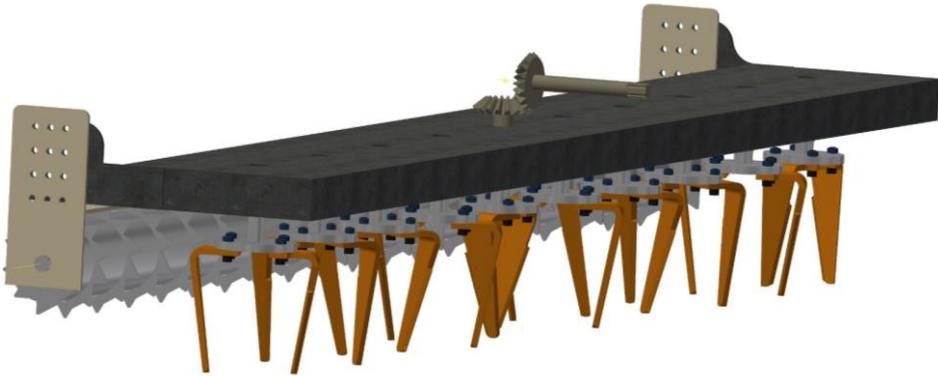
Figure 2: The simplified assembly model of the rotary harrow is based on Amazone KE3001 type machine [14]

The shapes of the harrow tines are extremely various. Most work machine manufacturers reduce fuel consumption and wear by using new, optimized shapes and innovative materials to increase the efficiency of their products [8].

The most commonly used harrow tine is the wide-bladed harrow tine. It can be made in straight or helical, vertical or oblique versions. In this paper we analyze a wide-bladed, oblique harrow tine. The width of the harrow tines increases from the tip to the base, so the cutting edge is able to pull the plant stalks into the soil, hence it has a slight compacting effect in the working depth and avoids any obstacles more easily.

2.2. The kinematic model

The trajectories of the points situated on the tines are important parts of the kinematic models. The trajectories can be used to identify parameters that affect tooth movement and to analyze how they affect the tillage process. The dimensions of soil chips formed during the movement of harrow tines in the soil can also be analyzed [10].

The points noted with $A_i(x_{Ai}, y_{Ai}, z_{Ai})$ and $B_i(x_{Bi}, y_{Bi}, z_{Bi})$, $i = 1 \div 10$ represent the tip of the harrow tines. The tines are moved by the combined effect of a rotating and a forward (towing) movement (*Fig. 3*).
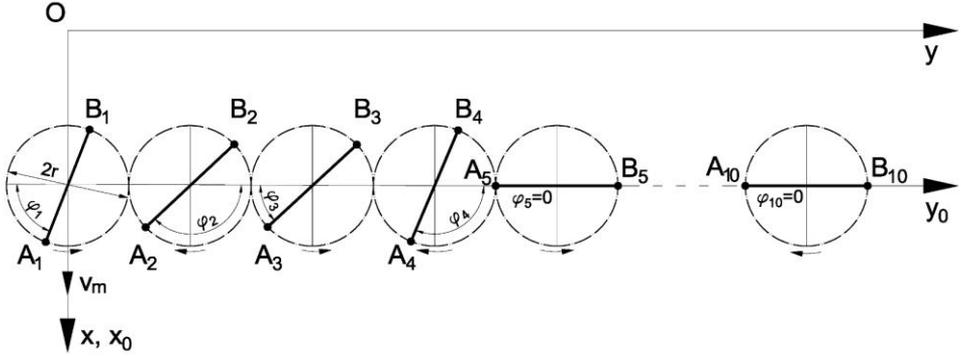
Figure 3: The initial positions of the working units, according to
Amazone KE 3001 [14]

The motion equations of the points $A_i(x_{Ai}, y_{Ai}, z_{Ai})$ and $B_i(x_{Bi}, y_{Bi}, z_{Bi})$ in the stationary coordinate system $°xO°y$ are given by the following relations:

$$°A_i \begin{cases} °x_{Ai} = r_t \sin\left((-1)^i\theta + \varphi_i\right) + v_m t \\ °y_{Ai} = 2r(i-1) + r_t \cos\left((-1)^i\theta + \varphi_i\right) \\ °z_{Ai} = -a \end{cases} \tag{1}$$

$$°B_i \begin{cases} °x_{Bi} = r_t \sin\left((-1)^i\theta + \varphi_i + \pi\right) + v_m t \\ °y_{Bi} = 2r(i-1) + r_t \cos\left((-1)^i\theta + \varphi_i + \pi\right) \\ °z_{Bi} = -a \end{cases} \tag{2}$$

$$\theta = \varpi t \tag{3}$$

where: v_m is the forward speed of the rotary harrow, [m/s]; ω is the angular velocity of the drive shaft, [rad/s]; a is the working depth [m]; i is the serial number of the implement; φ_i is the angle between the direction line of the working units and the horizontal line - the initial phase shift, r is the radius of the rolling circle of the gears, r_t is the half distance between the tips of the tines from a working unit.

Note: for i = 5 the above relations describe the driving unit; r_t depends on the shape of the tine: if $2r_t = 2r$, the tines are vertical, if $2r_t > 2r$ the tines are divergent and if $2r_t < 2r$ the tines are convergent (the tines are tilted).

The initial phase shifts of each working unit φ_i are given in *Table 1*:

Table 1: Initial phase shifts for each work unit

φ_1	φ_2	φ_3	φ_4	φ_5	φ_6	φ_7	φ_8	φ_9	φ_{10}
$\pi/3$	$5\pi/6$	$\pi/6$	$2\pi/3$	0	$2\pi/3$	$\pi/6$	$5\pi/6$	$\pi/3$	0

According to the motion equations, the trajectories of the tines describe looped cycloids (*Fig. 4*) [15]. The equations are considered valid and suitable to perform kinematic studies.

Figure 4: The rotating harrow tine tips trajectories

We examined the effect of changes of the tilt angle of the harrow tines, the towing speed, and the angular velocity on moving trajectories.

The effect of the tilt angle of the harrow tines in the cultivated soil, as a function of the working depth, can be seen in *Fig. 5*.

Figure 5: Trajectories described by points in different depths of the tine.
a.) top of harrow tine, b.) middle of harrow tine, c.) tip of the harrow tine

Due to the cohesiveness of the teeth, the distance between the sides of the teeth decreases from the surface of the soil to the tip of the teeth. Thus, soil shredding is expected to be more vigorous close to the surface relative to the working depth. Thus, the inclination angle contributes to the proper formation of the seedbed.

The effect of the forward speed on the worked surface can be seen in *Fig. 6*. As the forward speed increases, the stride of the implement increases, and the cuts become less frequent. As a result, the seedbed becomes clumpier and may require further processing.
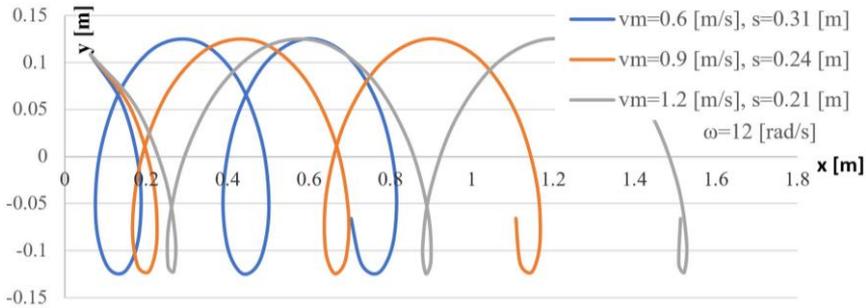
Figure 6: The effect of the towing speed on the trajectory of the tip

In practice as an illustrative indicator, the step of the implement [9], is used to characterize the kinematics of the work process. In the case of a rotary harrow, step *s*, the distance made during one revolution, can be calculated as follows:

$$s = v_m T = v_m \frac{2\pi}{\omega}; \qquad s = \frac{v_m}{n} \ [\text{m/rot}] \qquad (4)$$

where: T [s], period of rotating motion and n [rpm] the spindle speed.

It can be concluded, that the pitch of the rotary harrow varies in direct proportion to the forward speed and in inverse proportion to the speed.

The effect of the angular velocity on the worked surface can be seen in *Fig. 7*. A low angular velocity results in a less machined surface.
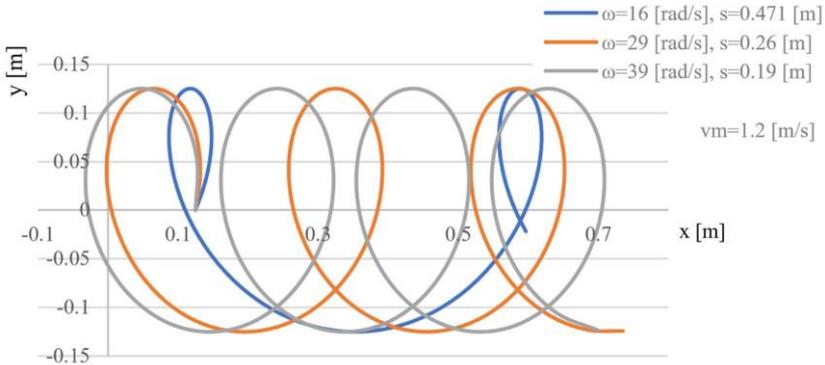
Figure 7: The variation of the tip trajectory for different angular velocities of the working unit

The rotary harrow's kinematic index is given by:

$$\lambda = \frac{v_p}{v_m} = \frac{\omega\, r}{v_m}\ [-] \tag{5}$$

Recommended values based on operating experience: $\lambda = 1.5 \div 6$, [11], [14]. In the investigated cases: $\lambda = 1.25 \div 2.50$, $\lambda = 1.66 \div 4.06$.

2.3. Dynamic modeling

In this paper, we consider the work of the tines divided into two elementary processes. The tines cut and compress the soil.

According to these processes in *Fig. 8*, the following forces are considered:

- the bit force (F_b);
- the friction forces (T_1 and T_2) - on the sides of the tine;
- the normal forces (N_1 and N_2) - acting on the tine sides;
- the compaction forces, F_c.

The forces acting on the tine were calculated based on literature data [16], [17], [18].

Determining the forces and the loads on the implements, which result from a real tillage of the soil, is a complex task, as the stresses already laying in the soil can also influence the soil's response to tillage forces, and these should be incorporated into the models, [19].

2.3.1. Determination of the Bit Force

The tine is considered a wedge-type blade having a pentagonal cross-section with parallel sides (*Fig. 8*) [16], [20], [21].

The tine, due to the work of the previous tine, also travels in shredded-loosened-pre-compacted soil, so the deformation zone will be asymmetrical. As a result of the change in the deformation zone, the forces acting on the tool also change. In determining the shear force, it is assumed that the physical properties of the soil are different on both sides of the harrow tooth, so the soil can be considered more compact on one side.

The forces acting on the wedge, in the light of the above, are shown in *Fig. 8*. The resulting force on the wedge is given by the sum of the normal forces N and the frictional forces T, acting on the wedge. The occasionally concentrated loads (given by stones, roots etc.) acting on the tool edge are not taken into consideration in our model.
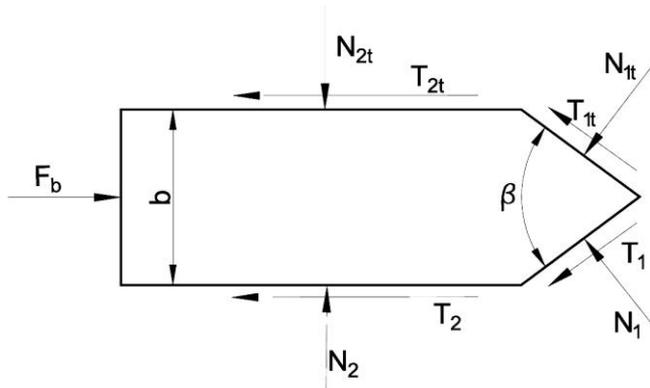
Figure 8: The forces acting in a section of a tine

The bit force acting on the tine can be calculated using the cutting resistance. It derives from normal forces and frictional forces, where:

$$N_{1t} = k_{1t}A_1 \text{ [N]}, N_1 = k_1A_1 \text{ [N]} \tag{6}$$

$$T_{1t} = \mu_t N_{1t} = \mu_t k_{1t}A_1 \text{ [N]}, \ T_1 = \mu N_1 = \mu k_1 A_1 \text{ [N]} \tag{7}$$

$$N_{2t} = k_{2t}A_2 \text{ [N]}, N_2 = k_2 A_2 \text{ [N]} \tag{8}$$

$$T_{2t} = \mu_t N_{2t} = \mu_t k_{2t}A_2 \text{ [N]}, \quad T_2 = \mu N_{2k} = \mu k_2 A_2 \text{ [N]} \tag{9}$$

$$F_b = N_1 \sin\frac{\beta}{2} + T_1 \cos\frac{\beta}{2} + T_2 + N_{1t}\sin\frac{\beta}{2} + T_{1t}\cos\frac{\beta}{2} + T_{2t} \text{ [N]} \tag{10}$$

where: k_{1t}, k_{2t}, are the specific resistances to worked soil deformation [N/m^2]; k_1, k_2 are the specific resistances to unworked soil deformation [N/m^2]; A_1 is the active surface of the tine edges [m^2]; A_2 is the surface of one of the sides of the

tine in contact with the soil [m^2]; μ_t, μ are the friction coefficients between the soil and the tine; and β is the lip angle of the tine.

During the work the A_1 and A_2 surface can be considered functions of the working depth a [m], $A_1 = f_1(a)$, $A_2 = f_2(a)$.

The k_1 and k_2 coefficients depend on: the soil deformation – are functions of thickness of the tine, the friction coefficient between the soil and the tine and the soil texture type. According to [12] $k_1, k_2 = f(b, \beta, \varphi)$, where b is the harrow tine thickness [m].

The values of the above coefficients are $\mu_t = 0.6$ and $\mu = 0.7$; $k_{1t} = 720883.68$ [N/m^2], $k_1 = 783930.62$ [N/m^2], $k_{2t} = 33692.35$ [N/m^2], $k_2 = 32128$ [N/m^2]; $\beta = 90°$ and $b = 0.015$ m.

2.3.2. Calculating the compaction forces, Fc

The harrow tines shred the soil, while they rotate during the work, compacting the soil in the towing direction. This process, shown in *Fig. 9*, can be considered a cutting with a variable wide profile (the projection of the width changes according to the rotation angle) [16]. The maximal width of the profile in this case is the length of the side of the tine.

To determine the compaction forces, the physical and mechanical properties of the soil, the geometry of the cutting tool must be considered [22].
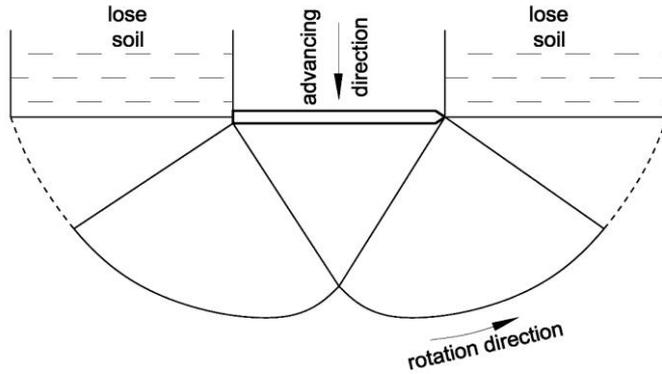
Figure 9: The compaction effect of the tine

To calculate soil resistance, the international literature uses the equation of Reece [18]. Cutting resistance per unit width:

$$f_c = \gamma a^2 N_\gamma + caN_c + a_{adh}aN_a + qaN_q \ \left[\frac{\text{N}}{\text{m}}\right] \qquad (11)$$

where: a is the working depth [m]; γ is the specific weight of soil [N/m^3]; c is the soil cohesion [N/m^2]; a_{adh} is the soil adhesion [N/m^2]; q surcharge on the soil surface [N/m^2]; N_c, N_γ, N_a, N_q are dimensionless Reece's resistance factors.

Reece's resistance factors are function of friction factors (ρ and φ), tool geometry, and deformation zone formation. Values used in the calculation: $\gamma =$ 1500·9.81=14715 [N/m^3]; $c = (0.01…0.02) \cdot 10^4$ [N/m^2]; $a_{adh} = 0$ [N/m^2]; $q = 0$ [N/m^2]; Reece's resistance factors: $N_c = 4.8$, $N_\gamma = 8.8$ ($\rho = \varphi = 35°$, $\alpha = 90°$ [16]).

The compaction force acting on a tine changes during its circular motion (*Fig.10*), and depends on the projected width of the harrow tine:

$$F_c = f_c \, l \, \cos\theta = f_c \, l \, \cos\omega t \text{ [N]} \tag{12}$$

where: l is the median width of the tine (at the half of the working depth a) [m].
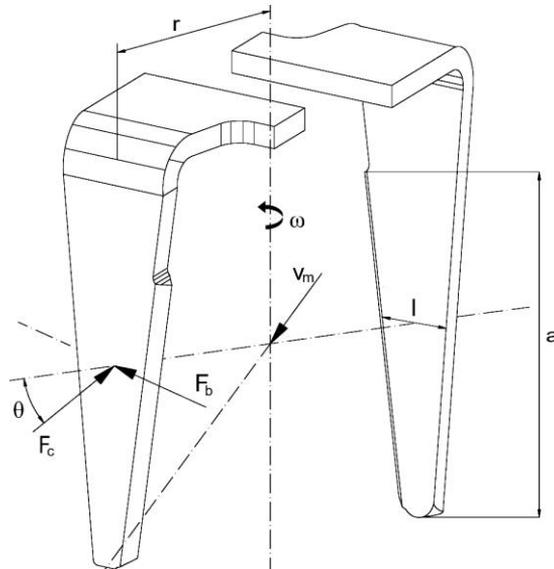
Figure 10: The position of the bit (F$_b$) and compacting (F$_c$) loads regarding the tines

2.4. Determining the driving torque

The paper [23] presents a methodology for mechanism simulation in case of agricultural machines, which may be applied at this mechanism also as seen in *Fig. 11*.
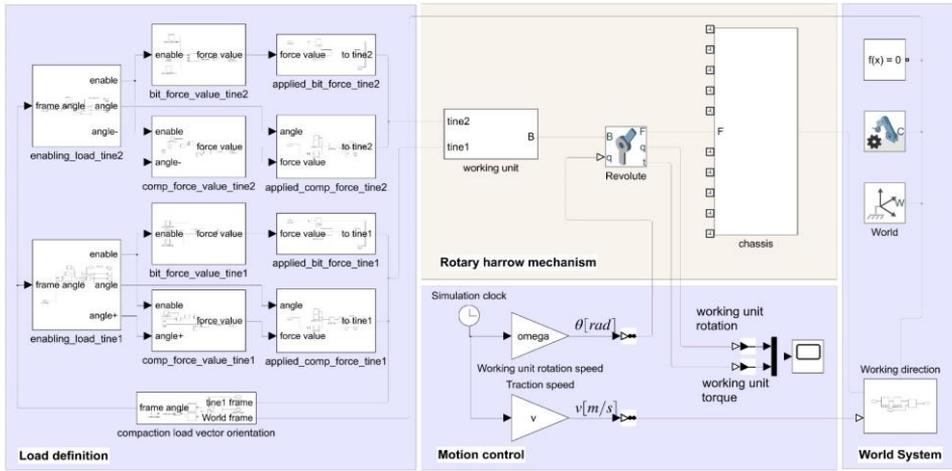
Figure 11: The Matlab Simscape model of the rotary harrow mechanism containing the working loads and kinematic parameters of the working process

The core element of the model is the rotary harrow mechanism area, which was presented already in *Fig. 2*. Because of model redundancy, the mechanism is simplified: besides the chassis, only one working unit is considered. This will allow a quicker simulation, and further, to obtain the driving torque for the whole rotary harrow machine, the number of the working units must multiply the simulation result. Overlapping the 10 torque variations through time must be done by taking into consideration the initial phase shifts depicted in *Table 1*.

Next to the *Rotary harrow mechanism* area, the *Motion control* allows to run the simulation with the kinematic parameters as *v* for the traction speed and *omega* for the working unit rotation speed. To obtain the needed driving torque for the working unit, the loads on the tines must be calculated and applied at the right timing intervals. As above presented (*Fig. 10.*), two loads were taken into account, whose definitions are provided in equations (10) and (12). The *Load definition* area from the simulation implements those equations and applies to the tines of the working unit (denoted by *tine1* and *tine2*). The magnitude and the direction of the loads are the same for both of tines, but the applying time interval differs through the simulation. Therefore, to have a more readable simulation model, the loads are calculated and presented separately.

The calculated driving torque for one working unit could be retrieved from the simulation through the graph element inside the *Motion control* area. Applying the phase shifts between the working units and summing the overall torques, for the parameters *v* = 1m/s traction speed and *omega* = 16 rad/s working unit rotation speed, the final simulation result is obtained.
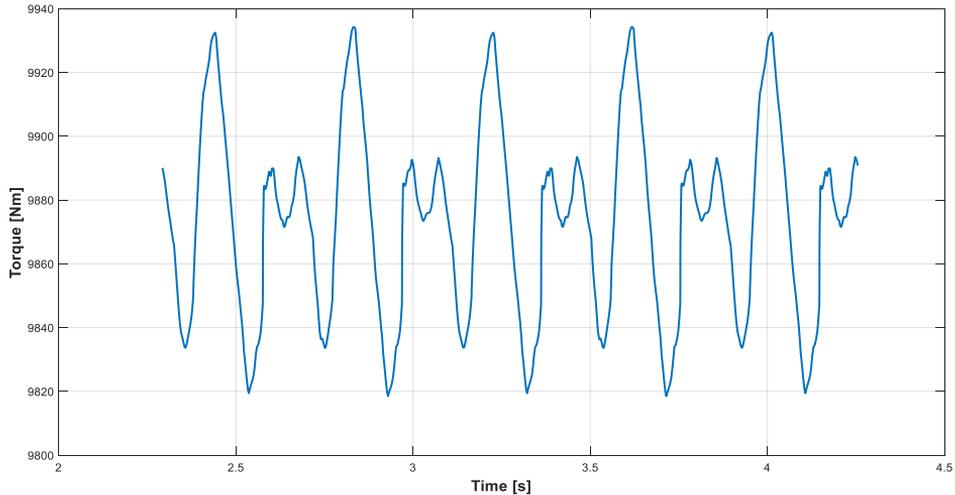
Figure 12: The variation of the driving torque (five cycles depicted)

Fig. 12. presents the torque variation through [2.25 s, 4.25 s] time interval consisting of five cycles. The mean value for the driving torque is 9872.6 Nm, having a cyclic variation in the [9818.5 Nm, 9934.2 Nm] interval.

3. Conclusions

The above defined kinematic equations describe trajectories also found in the literature, so they are suitable for performing theoretical investigations. As they are graphically visualized in a virtual environment, they easily allow to track changes made to various working parameters.

As the towing speed and the implement's spindle speed majorly affects the quality and the bumpiness of soil work to be performed, it is recommended to carefully choose the values for these parameters.

We can also state that the further development of soil preparation machines is inconceivable without a solid knowledge of the soils and their characteristics. The complex structure and inhomogeneity of soils make it very difficult to describe their general mechanical laws and to select the correct mechanical properties. The main difficulty in modeling a close to reality soil-machine relationship is to build a complex computer model of the soil. While tillage machines change the physical properties of the soil, they also have reactions on the implement. The currently used soil characteristics do not describe correctly the mechanical behavior of soils under all conditions. In the presented modeling process, we performed simplifications that take from the complexity of the tillage processes. The presented work can be used as base for further practical and experimental investigations.

References

[1] Láng, Z., "A zöldség-, dísznövény- és szaporítóanyag – termesztés berendezései és gépei", Publisher Mezőgazda, Budapest, 1999.

[2] Szendrő, P., "Géptan", Publisher Mezőgazda, Budapest, 2003.

[3] Pásztor, J., Forgó, Z., "Ásógép kinematikája és munkaminőségi mutatóinak vizsgálata növényházban", *Műszaki Szemle, Publisher EMT, Cluj-Napoca*, 2009.

[4] Birkás, M., Dekemati, I., Kende, Z., Radics, Z., Szemők, A., "A sokszántásos műveléstől a direktvetésig-Előrehaladás a talaj művelésében és védelmében", *Agrokémia és Talajtan*, no.67, vol. 2, pp.253−268, 2018.

[5] Raparelli, T., Pepe, G., Ivanov, A., Eula, G., "Kinematic Analysis of Rotary Harrows", *Journal of Agricultural Engineering*; vol. LI:976, pp. 9−14, 2020.

[6] Kovari, A., "Effect of Leakage in Electrohydraulic Servo Systems Based on Complex Nonlinear Mathematical Model and Experimental Results", *Acta Polytechnica Hungarica*, vol. 12, no. 3, pp. 129–146, 2015.

[7] Sánta, R., Garbai L., Fürstner, I., "Numerical Investigation of the Heat Pump System", *Journal of thermal analysis and calorimetry*, vol. 130, no 2, pp. 1133−1144, 2017.

[8] Balsari, P., et al., "Performance Analysis of a Tractor - Power Harrow System Under Different Working Conditions", *Biosystems Engineering*, vol. 202, pp. 28−41, 2021.

[9] Major, T., "Mechanization Development for Soil Tillage at Stumpy Forest Areas", *Theses of doctoral (PhD) dissertation, University of West Hungary, Sopron*, 2014, http://doktori.nyme.hu/458/2/ Tezisfuzet_angol_MajorT.pdf.

[10] Virág, S., "Effects of Cultivation on Clod Formation os Soils Energetic Relationships in Breakage of Clods", *Theses of doctoral (PhD) dissertation, University of Debrecen*, 2005, https://dea.lib.unideb.hu/dea/handle/2437/79555.

[11] Naghiu, A., Baraldi, G., Naghiu, L., "Mașini și instalații agricole", *Editura Risopront, Cluj-Napoca*, 2004.

[12] Kovari, A., "Influence of Cylinder Leakage on Dynamic Behavior of Electrohydraulic Servo System," in *SISY 2009 - 7th International Symposium on Intelligent Systems and Informatics*, 2009, pp. 375−379.

[13] Sánta, R., "Comparative Analysis of Heat Pump System with IHX using R1234yf and R134a", *Periodica Polytechnica Mechanical Engineering*.

[14] https://amazone.hu/hu-hu.

[15] Máté, M., "Műszaki mechanika – kinematika", *Publisher TMS, Cluj-Napoca*, 2010.

[16] Sitkei, Gy. "Soil Mechanics Problems of Agricultural Machines", *Publisher Franklin Book Programs, New York, U.S.A.*, 1976, pp. 22−62.

[17] Major, T., Csanády, V., "Application of Numerical Analysis for the Design of Rotating Tool", *Hungarian Agricultural Engineering*, no. 26/2014, pp. 16−19, 2014.

[18] Chung, S. O., Sudduth, K. A, "Soil Failure Models for Vertically Operating and Horizontally Operating Strength Sensors", *American Society of Agricultural and Biological Engineers*, vol. 49(4), pp.851−863, 2006.

[19] Rajaram, G. "Mechanical Behavior of an Agricultural Soil", *Retrospective Theses and Dissertations. 9570, Iowa State University, Capstones*, 1991.

[20] Amantayev, M., Gaifullin. G., Nukeshev, S., "Modelling of the Soil-Two Dimensional Shearing Tine Interaction", *Bulgarian Journal of Agricultural Science*, 23, no.5, pp. 882−885, 2017.

[21] Aluko, O.; Seig, D. A., "An Experimental Investigation of the Characteristics of and Conditions for Brittle Fracture in Two-Dimensional Soil Cutting", *Soil & Tillage Research*, vol. 57, no.3, pp.143−157, November 2000.

[22]   Thakur, T. C., Godwin, R. J., "The Present State of Force Prediction Models for Rotary Powered Tillage Tools", *Journal of Terramechanics*, vol. 26, no. 2, pp. 121-138, 1989.
[23]   Forgó, Z.; Tolvaly-Roșca, F.; Pásztor, J.; Kovari, A., "Energy Consumption Evaluation of Active Tillage Machines Using Dynamic Modelling", *Appl. Sci.,* no.11, 6240, 2021.

Peculiarities of the Grinding Process of a Gear Hob Helical Rake Face

Norbert HODGYAI[1], Márton MÁTÉ[2], Ferenc TOLVALY-ROȘCA [2], Mircea Viorel DRĂGOI[1]

[1] Transilvania University of Brașov, Faculty of Technological Engineering
and Industrial Management, Department of Manufacturing Engineering,
e-mail: hodgyai@ms.sapientia.ro, dragoi.m@unitbv.ro
[2] Sapientia Hungarian University of Transylvania, Cluj-Napoca,
Faculty of Technical and Human Sciences, Târgu Mureș,
Department of Mechanical Engineering,
e-mail: mmate@ms.sapientia.ro, tferi@ms.sapientia.ro

Abstract: This paper presents a study regarding the gear hob's rake face grinding possibilities and its consequences. A simple theoretical lined surface is considered. The mathematical model of the reciprocate meshing of surfaces was applied. It was proven that the proposed form of the rake face cannot be obtained because an undercut of inacceptable extent occurs. It is also proven and sustained by CAD modeling that using a simplified, flat grinding disk, the undercut is avoided, but the phenomenon of transection appears.

Keywords: gear hob, rake face, grinding wheel, profile error, modeling.

1. Introduction

The gear-hob is one of the most productive cutting tools used in cylindrical gear machining. As well as known, literature contains a very large number of papers dealing with the cutting process, the generating process and the build-up of this very efficient cutting tool.

It can be stated that between the profile correction of the first attempt gear-hobs till nowadays high-quality tools spread a long way marked by research and results of a large number of scientists. A large number of papers deal with different profile modifications, the sharp differentiation of roughing, semi-finishing and finishing gear-hobs and their profiles [1], [2]. Here it must be remarked that interests are focused on profile modifications targeting the improvement of the tooth dedendum stiffness, extending to the study of the

realization, involving the profile of the gear hob tooth relief face grinding wheel, its axial profile and also its dressing methods [3], [4].

Other research results cluster near the cutting process. One stream is the study of the influence of the cutting parameters (the chip thickness, the feed, the cutting speed, and cutting depth) on the durability of the gear-hob's edge. The other stream focuses on the influence of the chip forming process on the thermal phenomenon, with effect on the tool life. Here have to be mentioned the results published in [5], [6], [7].

The real form of the rake face is narrowly discussed. The principle of the correct meshing of this surface is mentioned in a few papers [1], [8].

Despite of this, the present paper is based on the hypothetic presumption that the rake face form has a significant influence of the cutting edge form and thus, on the precision of the meshed gear.

The problem starts with the theoretical definition of the rake face. In classical approach, admitting a zero-addendum rake angle value, this is a constant helix parameter cylindrical helicoid surface, whose generator is a straight segment that intersects the helix axis. But trying to compute the corresponding grinding wheel profile, the phenomenon of undercut occurs. Trying to eliminate the undercut by limiting the grinding wheel diameter, another phenomenon – the transection – appears. This phenomenon occurs when the real body of the generating tool, a grinding wheel for example, or the subspace generated by the cutting edge in its relative motion, intersects the surface obtained with the application of the meshing theory [9], [10], [11], [12], [13].

While the phenomenon of the undercut [9], [10], [11] can be described with a robust mathematical model, the phenomenon of the transection is discussed merely involving numerical, CAD sustained simulation [12], [13]. Here is also important to mention the method of the successive subtracting [14], [15].

According to all mentioned above, it must be admitted that precisely designing the form and the grinding wheel profile of the helical rake-face of the tool is not a trivial task, due to the undercut of the sharpening tool. The generation of the surface is theoretically simple, but the practical implementation is complicated due to the transection. The developed mathematical model points to the peculiarities of the involved phenomenon, and is presented as follows.

2. The theoretical equations of the helical rake face

To create the ideal rake surface, we consider the mathematical - geometric model shown in *Fig. 1*. The $OX_0Y_0Z_0$ coordinate system is a stationary coordinate system. The revolution axis of the hob coincides with the X_0 axis of the stationary coordinate system. To simplify the graphical representation, the gear-hob is replaced with a cylinder. The $O_2X_2Y_2Z_2$ coordinate system is the coordinate

system attached to the gear-hob, that rotates about the X_2 axis according to the helical rake face rotation direction. The X_2 axis is fixed to the hob, and coincides with the X_0 axis of the stationary coordinate system. A grinding wheel is used to generate the rake face. The coordinate system fixed to the grinding wheel is the $O_1X_1Y_1Z_1$ coordinate system. The axis of the grinding wheel should be theoretically perpendicular to the pitch screw line of the rake face. To obtain the ideal helical rake face, we studied several cases when the grinding tool's axis was tilted with the angles corresponding to the declination angles of the helices situated on the flute base, addendum, dedendum respectively the pitch cylinder.
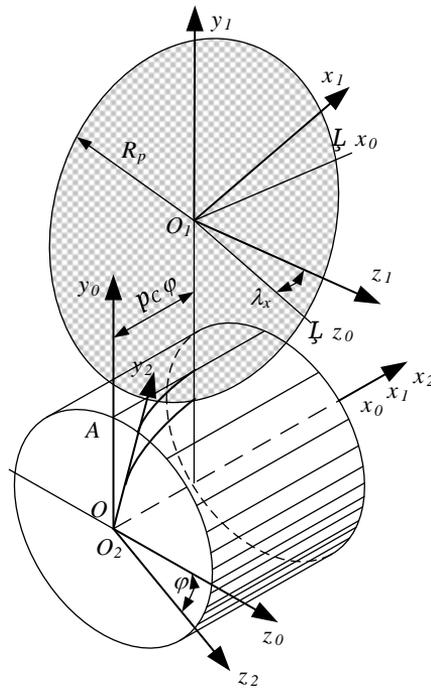
Figure 1: The considered coordinate systems

The equation of the rake face in matrix form [1], related to the frame S_2 of the gear hob, can be written as:

$$r_2(u, \varphi) = \begin{bmatrix} p\varphi \\ u\cos\varphi \\ -u\sin\varphi \end{bmatrix} \qquad (1)$$

where u is the radial and φ the angular parameter of the surface.

It is well known, that the normal vector of the ideal rake surface is defined as the vector product of the partial derivatives:

$$\underline{\boldsymbol{n}} = \dot{\boldsymbol{r}}_{2u} \times \dot{\boldsymbol{r}}_{2\varphi} \tag{2}$$

where the expressions of the operands involved in eq. (2) are:

$$\dot{\boldsymbol{r}}_{2u} = \begin{bmatrix} 0 \\ \cos\varphi \\ -\sin\varphi \end{bmatrix} \tag{3}$$

$$\dot{\boldsymbol{r}}_{2\varphi} = \begin{bmatrix} p \\ -u\sin\varphi \\ -u\cos\varphi \end{bmatrix} \tag{4}$$

Using the antisymmetric matrix of $\tilde{\dot{\boldsymbol{r}}}_{2u}$

$$\tilde{\boldsymbol{r}}_{2u} = \begin{bmatrix} 0 & u\cos\varphi & -u\sin\varphi \\ -u\cos\varphi & 0 & -p \\ u\sin\varphi & p & 0 \end{bmatrix} \tag{5}$$

and using (2), (4) and (5) the normal vector can be written as:

$$\underline{\boldsymbol{n}} = \tilde{\boldsymbol{r}}_{2u} \cdot \dot{\boldsymbol{r}}_{2\varphi} = \begin{bmatrix} 0 & u\cos\varphi & -u\sin\varphi \\ -u\cos\varphi & 0 & -p \\ u\sin\varphi & p & 0 \end{bmatrix} \begin{bmatrix} 0 \\ \cos\varphi \\ -\sin\varphi \end{bmatrix} = \begin{bmatrix} 0 \\ p\sin\varphi \\ p\cos\varphi \end{bmatrix} \tag{6}$$

On the other hand,, considering (1) and (6), the general parametric equation of the normal line in an arbitrary point of the helicoid (1) results in the following form:

$$\frac{x_2(u,\varphi) - p\varphi}{u} = \frac{y_2(u,\varphi) - u\cos\varphi}{p\sin\varphi} = \frac{z_2(u,\varphi) + u\sin\varphi}{p\cos\varphi} \tag{7}$$

The supposed contact curve between the grinding wheel surface and the helical rake face can be computed if a dependence between the independent surface parameters (u, φ) is found. Classically, this is obtained by the application of the kinematic theory of meshing, applied first time by Litvin [9]. A simpler way is based on the recognition that the normal line of a revolution surface intersects the surface axis. Thus, from all surface points must be selected only those where the surface normal intersects the axis of the grinding wheel. In algebraic terms, it can be primed that the linear system built up from the axis equations and the normal line equations must admit a unique solution.

Because the system comprises four linear equations and only 3 unknowns x_2, y_2, z_2, the sole characteristic determinant of the system must be zero.

Fig. 1 lets us observe easily that the revolution axis of the grinding wheel admits the following equations:

$$\begin{cases} y_2 = a_w \\ \tan \lambda_x = \dfrac{x_2}{z_2} \end{cases} \tag{8}$$

As it was stated earlier, the model deals with different tilting angles λ_x of the grinding wheel's axis, matching the declination angles of the helices situated on the characteristic diameters of the hob. On the other hand, a_w is the axial distance between the hob and the grinding tool.

Thus, using (8) and (7), the linear system becomes:

$$\begin{cases} y_2 = a_w \\ \tan \lambda_x = \dfrac{x_2}{z_2} \\ p \sin \varphi \, x_2 - u y_2 = p^2 \varphi \sin \varphi - u^2 \cos \varphi = e_1 \\ p \cos \varphi \, y_2 - p \sin \varphi \, z_2 = pu = e_2 \end{cases} \tag{9}$$

The characteristic determinant of the system is to be written as:

$$\Delta_c = \det \begin{bmatrix} 0 & 1 & 0 & a_w \\ 1 & 0 & -\tan \lambda_x & 0 \\ p \sin \varphi & -u & 0 & e_1 \\ 0 & p \cos \varphi & -p \sin \varphi & e_2 \end{bmatrix} \tag{10}$$

Equalizing this to zero and doing the calculus, it results the following dependence between the helicoid parameters:

$$u^2 \cos \varphi - (a_w + p \tan \lambda_x) u + a_w p \varphi \tan \lambda_x - p^2 \varphi \sin \varphi = 0 \tag{11}$$

Due to the fact the condition above is an algebraic equation of 2nd degree in u, there exist two solutions:

$$u_{1,2} = \frac{a_w + p \tan \lambda_x \pm \sqrt{(a_w + p \tan \lambda_x)^2 - 4(a_w p \varphi \tan \lambda_x - p^2 \varphi \sin \varphi) \cos \varphi}}{2 \cos \varphi} \tag{12}$$

In order to decide which solution corresponds to the geometric reality of the grinding operation, a numerical approach is necessary. Let's consider the case of a gear-hob originating from an involute worm. Starting from the normal module, the pitch helix angle and the normal rake profile angle, all the other characteristic parameters were computed using the classical formulae from the literature [8], [9]. Omitting the computation, the resulting geometrical data are the following:

The input data for the calculation are: normal module m_n: 5   [mm],   pitch helix declination angle λ_0= 5 [deg], rake profile normal angle α_{0n}= 20 [deg].

Omitting the computation, the calculated geometrical data are the following: frontal module m_t= 57,36587 [mm], rake profile frontal angle α_{0t}= 76°32'1", pitch radius R_0 = 28.684 [mm], basic cylinder radius R_b = 6.68 [mm], addendum cylinder radius R_a = 34.934 [mm], dedendum cylinder radius R_f = 22.434 [mm], flute interior radius R_q= 14.984 [mm], flute interior helix declination angle λ_q=

2.617 [deg], dedendum helix declination angle $\lambda_f = 3.914$ [deg], addendum helix declination angle $\lambda_a = 6.082$ [deg], worm helix parameter $p = 2,50955$ [mm], flute helix parameter $p_C = 327.8628563$ [mm].

Now, computing both $u(\varphi)$ dependencies given by (12) and associating them the graphical representation shown in *Fig. 2*, it can be decided that the real considered situation corresponds to the concave line, because the u values involved match the interval $R_q \leq u \leq R_a$. As a conclusion, the applicable $u(\varphi)$ dependence is:

$$u_{1,2} = \frac{a_w + p \tan\lambda_x - \sqrt{(a_w + p \tan\lambda_x)^2 - 4(a_w p\varphi \tan\lambda_x - p^2\varphi \sin\varphi)\cos\varphi}}{2\cos\varphi} \quad (13)$$

Representing function (13) for different helix angles values, it can be concluded that the maximum point of the graphic increasingly sharpens. Thus, acceptable u parameter values occur for more and more narrowing intervals of φ, as shown in *Fig. 3*. Inspecting the figure, it results that even the accepted solution contains two possibilities for constituting the $u(\varphi)$ dependence. The first one results for the negative, while the other for the positive φ values.
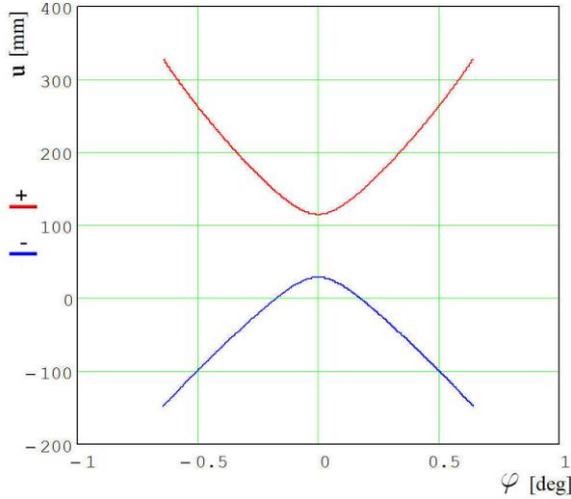
Figure 2: The graphics associated with function $u(\varphi)$
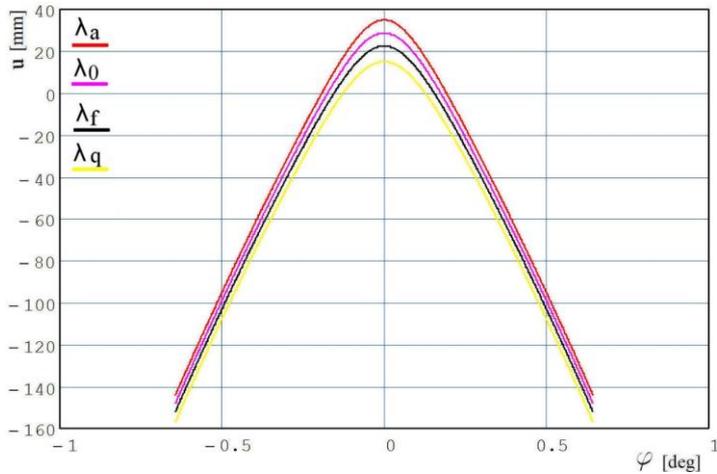
Figure 3: The variation of the accepted $u(\varphi)$ function in function of the helix angles

Next goal is to build up the contact curves and using them, the grinding wheel surface. The model was numerically solved. Using the Mathcad environment, equation (11) was numerically solved for a number of $n = 50$ discrete and equidistant u values, $u \in [R_q, R_a]$, searching first the negative and after that the positive solutions in variable φ. The procedure was repeated for the next values of the grinding wheel axis declination angle: $\lambda \in \{\lambda_q, \lambda_f, \lambda_0, \lambda_a\}$.
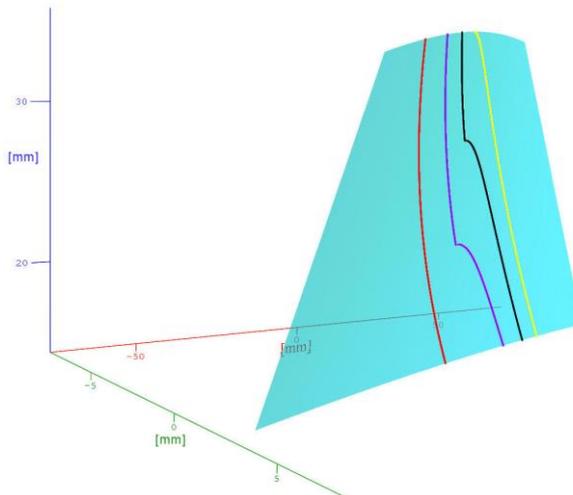
Figure 4: Contact curves for the positive φ values
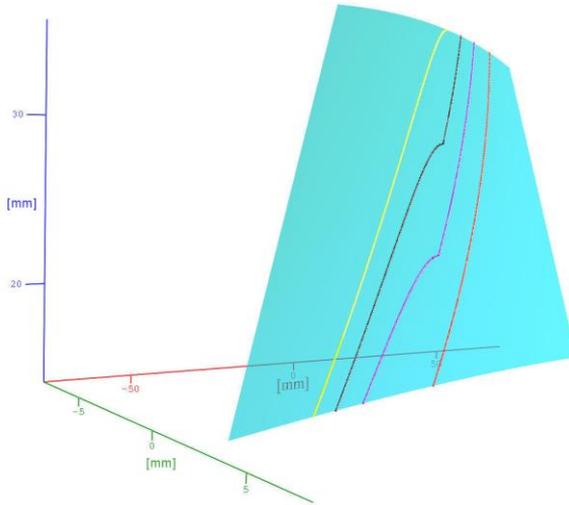
Figure 5: Contact curves for the negative φ values

The possible contact curves are represented through their discrete points, replacing the discrete arranged pairs. The contact curves on the theoretical helical rake face are shown in *Fig. 4* for the negative, respectively in *Fig. 5* for the positive φ values.

The colors used are red for the addendum, magenta for the pitch, black for the dedendum and finally yellow for the flute base helix angles.

Let's denote $\mathbf{C} = \left(\mathbf{x}^{(2)}, \mathbf{y}^{(2)}, \mathbf{z}^{(2)}\right)$ the matrix of the coordinates of the contact curve points, where $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are the columns of length n of the coordinates.

The axial profile of the grinding wheel can be computed by executing several computation tasks. First of all, the coordinates of the contact curve must be transposed in the frame of the grinding wheel. Here can be used the transformation matrix between frame S_1 and S_2, particularized for the $\varphi = 0$ parameter value. Thus, the contact curve parameters will be comprised in the matrix $C = \left(x^{(1)}, y^{(1)}, z^{(1)}\right)$. We omit here the detailed computation. Using the coordinate values in the frame of the grinding wheel, for each $C_i\left(x_i^{(1)}, y_i^{(1)}, z_i^{(1)}\right)$ corresponds a profile point $P_i\left(x_i^P, y_i^P, z_i^P\right)$ with the coordinates computed as follows:

$$x_i^P = \sqrt{\left(x_i^{(1)}\right)^2 + \left(y_i^{(1)}\right)^2}$$
$$z_i^P = z_i^{(1)} \tag{14}$$

Using (14), the revolution surface of the grinding wheel results immediately.

The relative position of the grinding wheel's surface and the theoretical rake face is shown in *Fig. 6*.
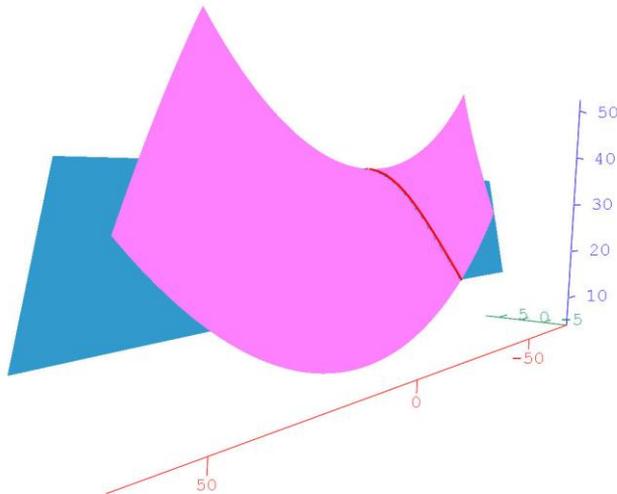
Figure 6: The relative position of the surfaces of the grinding wheel and the rake face

In *Fig. 6* the blue surface represents the ideal helical rake face and the pink one the side surface of the grinding wheel.

Based also on the figure, it can be concluded that in the position of the grinding wheel where the contact curve is correct, the grinding wheel cuts under the dedendum cylinder, i.e., it cuts into the body of the gear-hob.

As the results of the numerical analysis shows, it can be concluded that the diameter of the grinding wheel required to generate the contact curve is not suitable for grinding the rake face, because the disk cuts very deep into the hob body.

3. The practical generation of the helical rake face

Regarding the industrial generation of the gear-hob rake face we can consider two solutions: using a grinder wheel (or mill) or a grinding bit (or an end mill).

In the present paper we studied the finishing of the helical rake face of parametric equations (1), with a profiled grinding- and with a cylindrical grinding wheel.

3.1 Generating the rake face using a profiled grinding wheel

As it can be seen in *Fig. 7*, the axis of the grinding wheel is parallel to the X_0OZ_0 plane. The normal vector of the helical rake face passes through the axis of rotation of the grinding wheel.
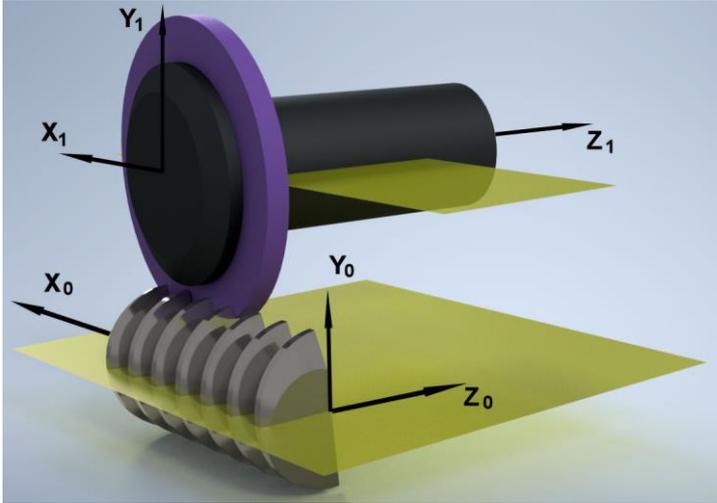
Figure 7: Rake face generation with a profiled grinding wheel

The location of the contact curves on the ideal rake face is in the negative angular range of φ (*Fig. 5*).

In order to be able to sharpen the gear-hob properly, the revolution surface of the grinding must not intersect the helical rake face.

3.2 Generating the rake face using a cylindrical grinding wheel

As a simplified procedure in some industrial practice, the side face (a plane surface) of a grinding wheel is used as the generator of the helical rake face. However, in the case of this simplified, practical solution, applying the appropriate equations, we do not obtain a solution to the applied function, from here we can conclude that there is no common normal, similarly to the previous cases for the undercut situation. The traces left by the flat disk are shown in *Fig. 8*.
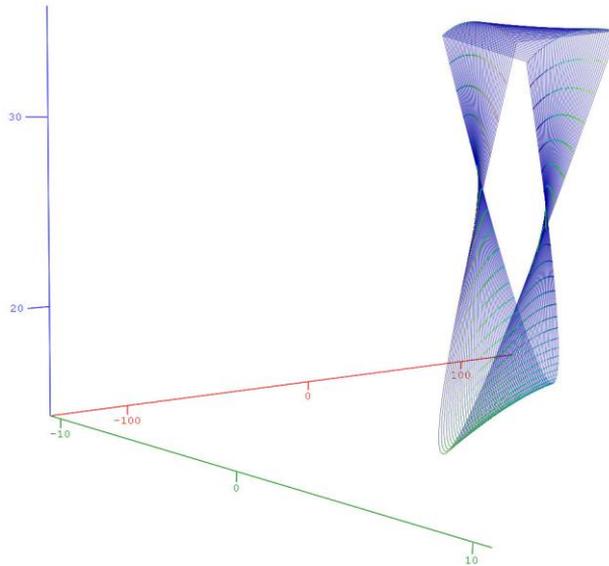
Figure 8: Numerical solution for the traces of a cylindrical grinding wheel

Generating the rake face with a cylindrical grinding wheel, and using the "Analyze Interference" command on the assembly model (*Fig.* 9) it can be clearly seen, that the grinding wheel undercuts the helical rake face in all cases.
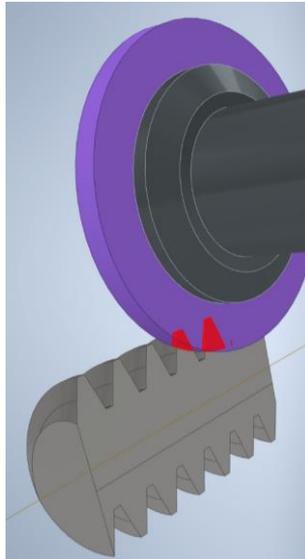
Figure 9: Analyze Interference command running on the assembly model

4. Conclusions

The revolution surface of the grinding tool was computed using the theory of meshing, but it was demonstrated that this firmly undercuts the machined rake face, even in cases when the contact is correct along the computed contact curve.

The flat grinding surface (a circular disk) produces through the relative motion reported to the gear hob (or to the surface) a points cloud that includes the theoretical surface subjected to the machining process. The theoretical surface cannot result because the phenomenon of transection appears in every region of action of the grinding wheel's generating surface. Here the undercut is surprisingly avoided.

The volume of the points cloud increases with the tilting angle of the grinding tool's axis, e.g. the phenomenon of transection appears with ascending intensity.

Considering the computing and the models presented it can be concluded that the theoretical form of the rake face (1) cannot be achieved through classical helical grinding operation, using disk-type tools.

If for economic reasons the classical grinding procedure is still applied, the computation of the edges and of the relief face must take into consideration the real form of the rake face, that results as the trace of transection.

Acknowledgement

References

[1]   Radzevich, S.P., "Gear Cutting Tools: Fundamentals of Design and Computation", *CRC Press*, NY, 2010.

[2]   Máté, M., "Hengeres fogaskerekek gyártószerszámai", *Erdélyi Múzeum-Egyesület*, *Kolozsvár*, 2016.

[3]   He, R., Tenberge, P, Xu, X., Li, H., Uelpenich, R., Dong, P., Wang, S., "Study on the Optimum Standard Parameters of Hob Optimization for Reducing Gear Tooth Root Stress", *Mechanism and Machine Theory*, vol. 156, 2021.

[4]   Yang, S., Chen, W., "Modeling and Experiment of Grinding Wheel Axial Profiles Based on Gear Hobs", *Chinese Journal of Aeronautics*, vol. 34, Issue 6, pp. 141−150, 2021.

[5]   Guba, N., Hüsemann, T., Karpuschewski, B., "Influence of Gear Hobbing Feed Marks on the Resulting Gear Quality After Discontinuous Profile Grinding", *CIRP Journal of Manufacturing Science and Technology*, vol. 31, pp. 314−321, 2020.

[6]   Karpuschewski, B., Beutner, M., Köchig, M., Härtling, C., "Influence of the Tool Profile on the Wear Behaviour in Gear Hobbing", *CIRP Journal of Manufacturing Science and Technology*, vol. 18, pp. 128−134, 2017.

[7]   Troß, N., Brimmers, J., Bergs, T., "Tool Wear in Dry Gear Hobbing of 20MnCr5 Case-Hardening Steel, 42CrMo4 Tempered Steel and EN-GJS-700-2 Cast Iron", *Wear*, 2021.

[8]   Hollanda, D. "Aschiere si scule", *Reprografia I.I.S., Tg. Mures*, 1982.

[9]   Litvin, F. L., Fuentes, A., "Gear Geometry and Applied Theory", *Cambridge University Press*, 2004.

[10] Dudás, I. "The Theory and Practice of Worm Gear Drives", *Penton Press, London,* 2000.

[11] Dudás, I. "Possibilities of Exact Grinding of Conical and Globoid Worms", *International Review of Mechanical Engineering (IREME)*, vol. 1., no. 3, pp. 200−207.

[12] Balajti, Z., Mándy, Z., "Proposed Solution to Eliminate Pitch Fluctuation in Case of Conical Screw Surface Machining by Apex Adjustment", *Procedia Engineering*, vol. 55, pp. 266−273, 2021.

[13] Balajti, Z., Ábel, J., Dudás, I., "Examination for Post-Sharpening Adjustment of Cutting Edge of a Worm Gear Hob with Circle Arched Profile in Axial Section", *Procedia Engineering*, vol. 55, pp. 260−265, 2021.

[14] Tolvaly-Roșca, F., Máté, M., Forgó, Z., Kakucs, A., "Development of Helical Teethed Involute Gear Meshed with a Multi-Edge Cutting Tool Using a Mixed Gear Teeth Modeling Method", *Procedia Engineering*, vol. 181, pp. 153−158, 2017.

[15] Tolvaly-Roșca, F., Forgó, Z., "Mixed CAD Method to Develop Gear Surfaces Using the Relative Cutting Movements and NURBS Surfaces", *Procedia Technology*, vol. 19, pp. 20−27, 2015.

Implementation of Grid Synchronization Methods on a Real Time Development System

Andor-Attila ZSIGMOND[1], András KELEMEN[2]

Sapientia-Hungarian University of Transylvania, Cluj-Napoca,
Faculty of Technical and Human Sciences, Târgu Mureș,
Department of Electrical Engineering,
e-mail: [1]zsigmonda@student.ms.sapientia.ro, [2]kandras@ms.sapientia.ro

Abstract: The paper presents a study regarding the implementation of two representative grid synchronization methods, intended to be used in the control structure of a three phase switch mode voltage rectifier.

The methods considered are the Synchronous Reference Frame Phase Locked Loop (SRF PLL) and the Double Synchronous Frame Phase Locked Loop (DSRF PLL), respectively. These synchronization methods have been compared on a real time development system type dSpace 1104 from the point of view of their performance in case of an unbalanced mains voltage system. Finally, the influence of the chosen grid synchronization method on the performance of a synchronous reference frame controlled three-phase switch mode voltage rectifier is studied by simulation, and better results are demonstrated to be provided by the use of the DSRF PLL.

Keywords: PWM rectifier, Phase Locked Loop, synchronous reference frame control, grid synchronization.

1. Introduction

The paper deals with the grid synchronization, which is an important task in the control of grid connected switch mode rectifiers [1], [9]. The voltage oriented control of the PWM voltage rectifiers is based on the identification of a reference frame synchronously rotating with the voltage vector of the direct sequence component of the three phase grid voltage system [2]. After the presentation of the operating principle of the SRF PLL and of the DSRF PLL [5], [6], the study compares the performance of these methods under unbalanced grid voltage conditions, in the simulated presence of a large inverse sequence component. Real-time models of the rectifier control structures including the two PLL versions are built using the Real Time Toolbox of Matlab-Simulink,

and are implemented on a dSpace 1104 type development system [8]. The full rectifier control structure is included to make possible the determination of the minimum program cycle. Practical measurements are made to compare the performance of the PLL structures in these conditions, for the secondary voltage system of a transformer with unbalanced load. Finally, the performance of the PWM rectifier is analyzed by discrete time simulation, using the maximum sampling frequency determined for the real time development system, in the case of the SRF PLL and DSRF PLL synchronization methods, under unbalanced grid conditions.

2. Grid synchronization methods

The goal is the identification of the angular position, magnitude and rotation frequency of the grid voltage vector. Out of the several existing phase locked loop versions [12], designed to perform this task, a well-known initial one is the SRF-PLL [4], [6], [7]. Its operation is based on the orientation of a coordinate system so that one of the grid voltage vector components in this coordinate system is cancelled. Thus, the cancelled component plays the role of the phase error, and the controller that cancels this error operates in this coordinate system, said synchronous, because in steady state it rotates in phase, and with the same frequency with the grid voltage vector. A block diagram is shown in *Fig. 1*, where θ_e is the estimated phase angle in the stationary coordinate system of the space vector assigned to the grid voltage system $[u_a, u_b, u_c]^T$, $[A]$ is the matrix of the $abc/\alpha\beta$ Clarke transformation, $[D(\theta_e)]$ is the matrix of rotation by θ_e, while ω_{ff} is the feedforward value of the angular frequency.
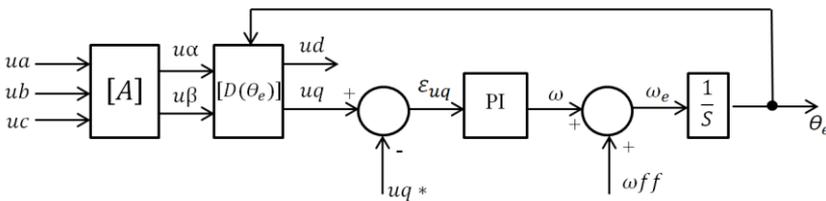
Figure 1: Block diagram of the SRF-PLL [4]

The main drawback of this PLL structure is its poor performance under unbalanced or distorted grid voltage [6], [14]. In this case, the essential task is the identification of the fundamental frequency positive sequence voltage component. The Matlab-Simulink simulation results shown in *Fig. 2* illustrate the inaccurate detection of the direct component of a 50 Hz, 3-phase unbalanced

voltage system, with components given in *Table 1*, in case of the controller parameters $K_p = 13.06$, $K_i = 1451$, $\omega_0 = 2\pi 40 \frac{rad}{s}$.

Table 1: Components of the voltage system from *Figure 2*

Component	Amplitude[V]	Initial phase[deg]
U_s^+	17	0
U_s^-	3.4	45
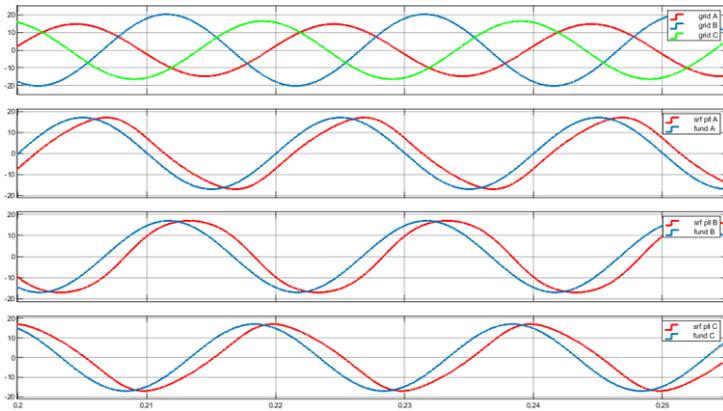
Figure 2: Unbalanced 3-phase grid voltage system (top) and its direct sequence component (blue waveforms on the other axes) vs. the 3-phase system that results using the phase angle detected with SRF-PLL
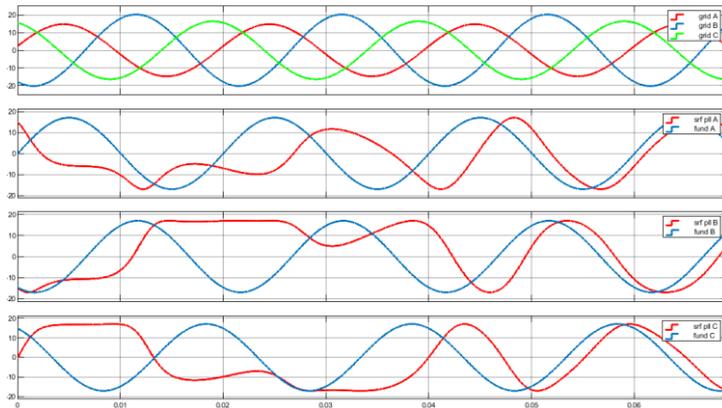
Figure 3: The starting transients of the SRF-PLL in case of unbalanced grid voltage

On the contrary, the Double Synchronous Reference Frame PLL method can successfully separate the direct and inverse components of the unbalanced grid

voltage, and can also deal with a distorted voltage system [2], [5]. This is performed by the PLL structure shown in *Figure 4,* using two synchronous reference frames with direct and inverse rotation senses, respectively. A decoupling network (*Figure 5*) cancels the double-frequency components that result from the representation of the direct and inverse sequence voltage vectors in the reference frames rotating in the opposite direction. Thus the low-pass filtering tasks of the PLL become easier and the dynamic response becomes better [5].

This structure can even be generalized for the detection of higher order harmonics [5].
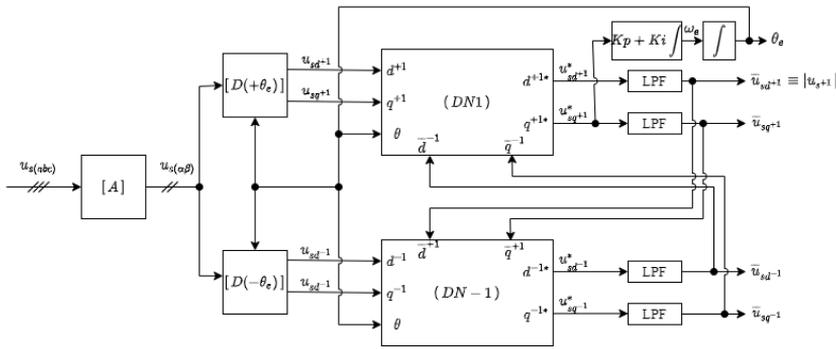
Figure 4: The block diagram of the DSRF-PLL [5]
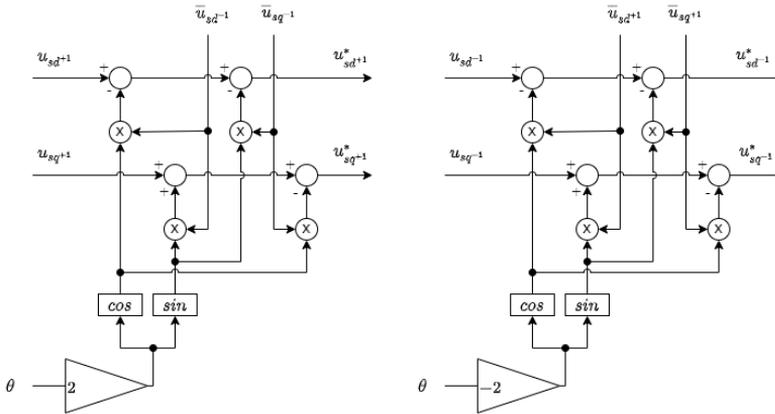
Figure 5: The block diagrams of the positive and negative sequence decoupling cells [5]

The estimated angular positions of the reference frames with $(+\omega)$ direct and $(-\omega)$ inverse angular speed are θ_e and $-\theta_e$, respectively.

In a similar way to the SRF-PLL case, *Figure 6* compares the direct sequence component of the voltage system to the voltage system detected by the DSRF phase locked loop, for the controller parameters $K_p = 13.06$, $K_i = 1451$ The initial value of the angular frequency was $\omega_0 = 2\pi 40 \ \frac{rad}{s}$, while the cut-off angular frequency of the low-pass filters was $\omega_f = 222.14 \ \frac{rad}{s}$.
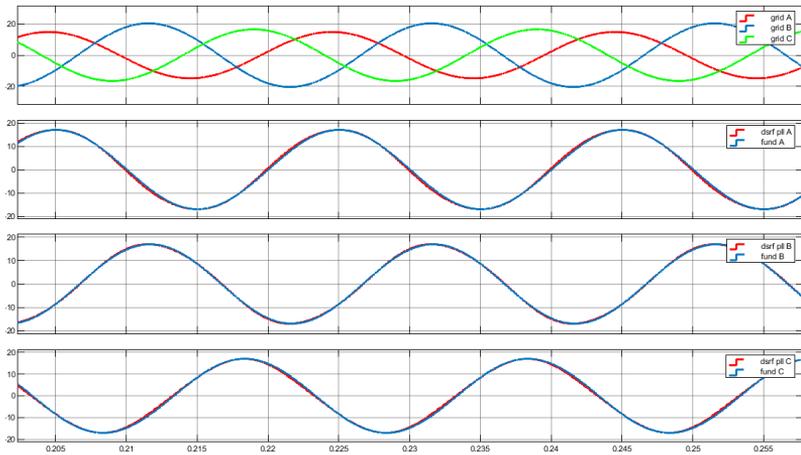
Figure 6: Unbalanced 3-phase grid voltage system (top) and its direct sequence component (blue waveforms on the other axes) vs. the 3-phase system that results using the phase angle detected with DSRF-PLL
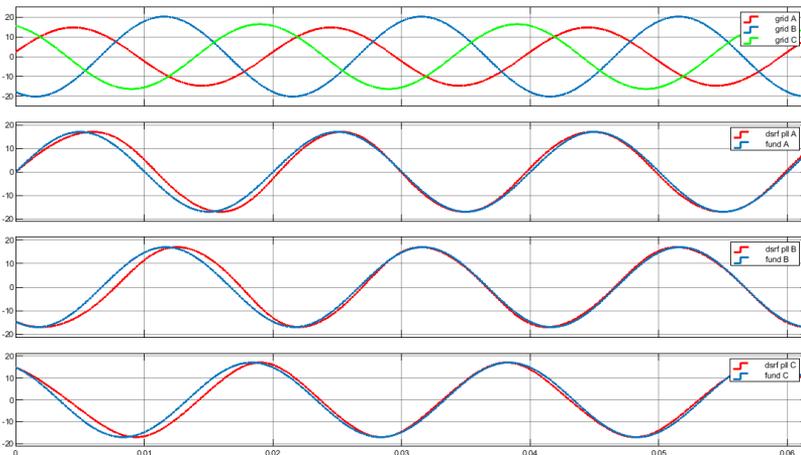
Figure 7: The starting transients of the DSRF-PLL in case of unbalanced grid voltage

3. The control structure of the switch mode voltage rectifier

The power circuit of the investigated switch mode voltage rectifier is shown in *Figure 8*.

The control structure is based on the mathematical model (1), developed in the synchronous reference frame [9], [11], [13].
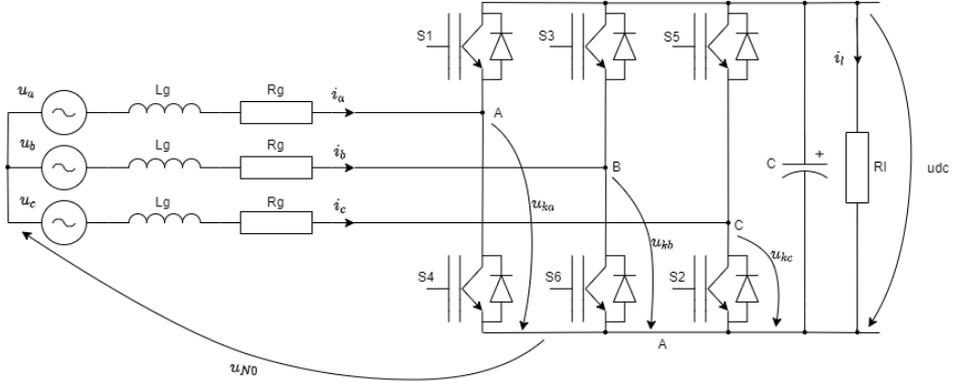
Figure 8: The power circuit of the 3 phase switch mode voltage rectifier [9]

$$\begin{cases} u_{k_d} = -L\dfrac{di_d}{dt} - Ri_d + \omega Li_q + u_d \\[2mm] u_{k_q} = -L\dfrac{di_q}{dt} - Ri_q - \omega Li_d + u_q \\[2mm] \quad C\dfrac{du_{dc}}{dt} = S_d i_d + S_q i_q - i_{load} \end{cases} \qquad (1)$$

where

$u_d, u_q, i_d, i_q, u_{k_d}, u_{k_q}$ are the d and q components in the synchronous reference frame of the grid phase voltage, grid current, and rectifier bridge input voltage, respectively.

$S_{d,q}$ are the components of the switching vector, defined by the switching states of the upper transistors.

The synchronous coordinate system voltage mode controller synthesizes the rectifier input voltage references using the compensation network from *Figure 9* [9]. This structure is based on (1), where

$$\Delta u_d = -L\frac{di_d}{dt} - Ri_d$$

$$\Delta u_q = -L\frac{di_q}{dt} - Ri_q \qquad (2)$$

are the outputs of the d and q axes current controllers.
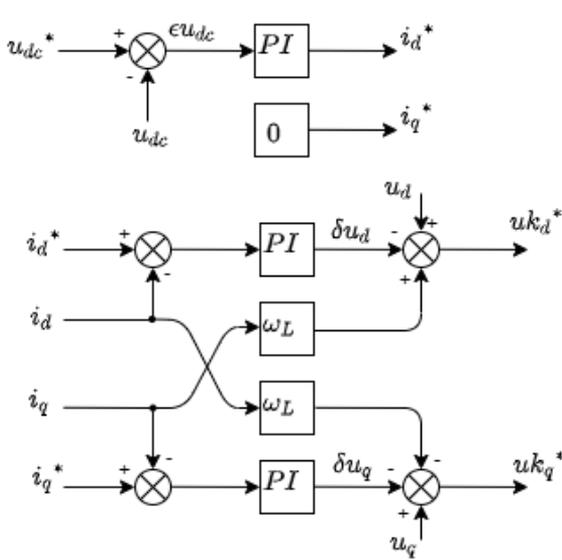
Figure 9: Control system in the synchronous reference frame [9]

The components of the grid current reference in the synchronous reference frame result from the assumption that the rectifier is operated at unity power factor. This means that the reference of the active grid current component i_d^* is delivered by the output of the DC-link voltage regulator, and $i_q^* = 0$ [15].

In consequence, the block diagram of the voltage controlled rectifier results as shown in *Figure 10*, where the rectifier input voltages result by space phasor modulation [11], based on the voltage reference components and the actual DC-link voltage.

The grid synchronization is performed by one of the above mentioned PLL structures, with significant influence on the performance of the rectifier, as discussed in the following sections.
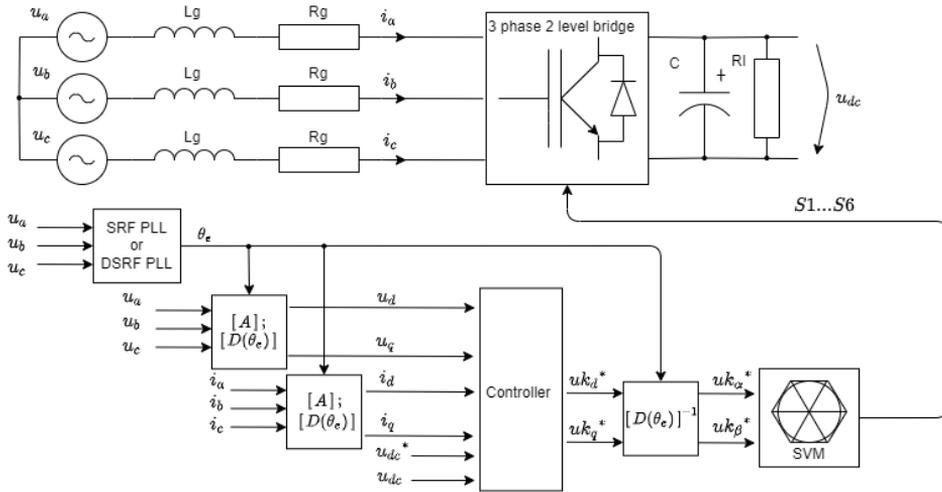
Figure 10: The overall control system of the switch mode rectifier with space vector modulated gate signals (SVM)

4. Experimental setup for the PLL assessment

The experimental setup for the analysis of the PLL structures is based on a dSpace DS1104 type real time development system (250 MHz floating point 603 PowerPC processor) [8], with galvanically isolated analog and digital inputs and outputs. The voltage source is the 17 V secondary of a 3-phase transformer, which delivers an unbalanced voltage system when heavily loaded.

One of the goals of this study has been the assessment of the maximum sampling frequency when all functionalities of the control system are implemented in the DS1104 system. The model from *Figure 11,* is used for the measurement of the execution times that finally limit the sampling frequency.
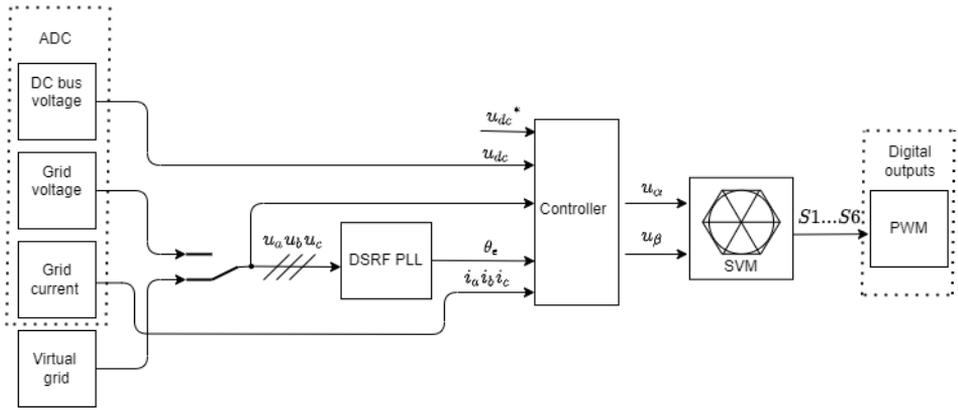
Figure 11: Real time model of the rectifier control system

Figure 12 shows the turnaround time corresponding to the case when all control functions were confined to a single task.
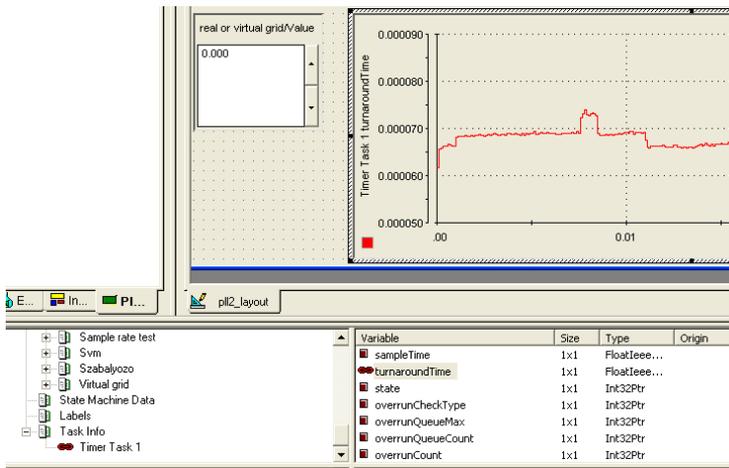
Figure 12: Timer Task 1 turnaround time

It can be observed that the turnaround time is less than $75 \, \mu s$, i.e. 75% of the initially set $100 \, \mu s$ cycle period maximum, which makes it possible to increase the sampling frequency to 12kHz, corresponding to an $83 \, \mu s$, sampling period.

It has to be mentioned that the sampling period can be further decreased by creating a multirate model, which allows for the execution of lower priority, but time consuming tasks in the time available between the processing of high-priority tasks [8].

The SRF and DSRF PLL synchronization methods have been compared while setting the above established maximum sampling frequency. The controller parameters have been set to the values given in the previous section.

In *Figure 13* there are represented on the same time diagrams the balanced grid voltage system and the voltage system identified by the corresponding PLL method. The bottom diagram shows the u_q outputs (representing in fact the phase error) and the angular positions identified by both methods. It can be observed that in case of a balanced grid voltage system, the behavior of both methods is similar. The small overshot in case of the DSRF-PLL is attributable to the controller settings, but the steady-state phase error is zero in both cases.
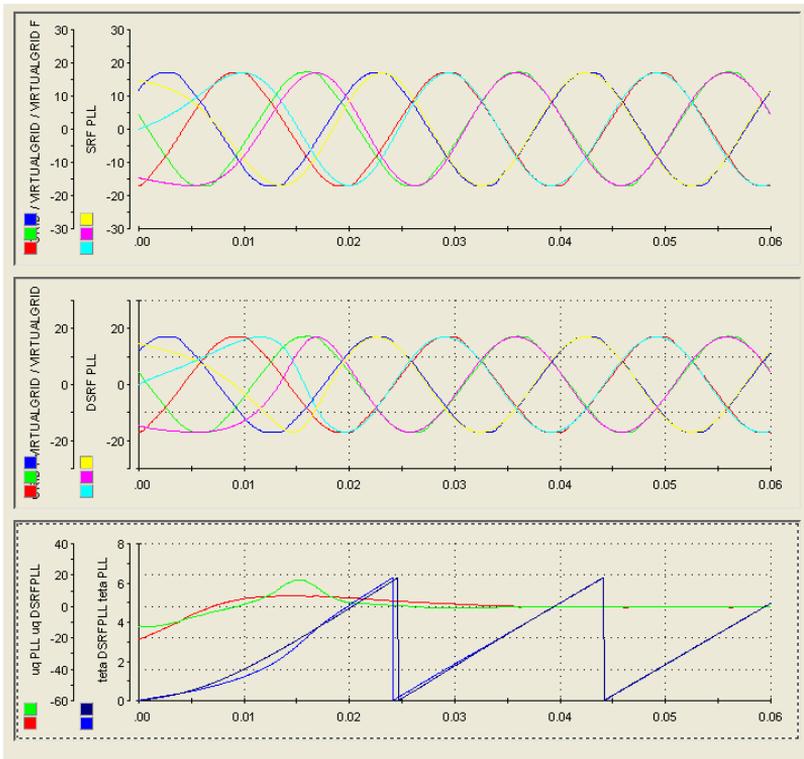
Figure 13: The starting transients while using for grid synchronization the SRF (top) and DSRF PLL (center) methods, in case of balanced grid voltages

Figure 14 presents the results of a start-up in case of an unbalanced grid simulated by the real time development system. In this case the amplitude ratio of the inverse and direct sequence components was 0.5. The steady-state error

of the DSRF-PLL method is zero, while the result obtained using the SRF-PLL method is unreliable.
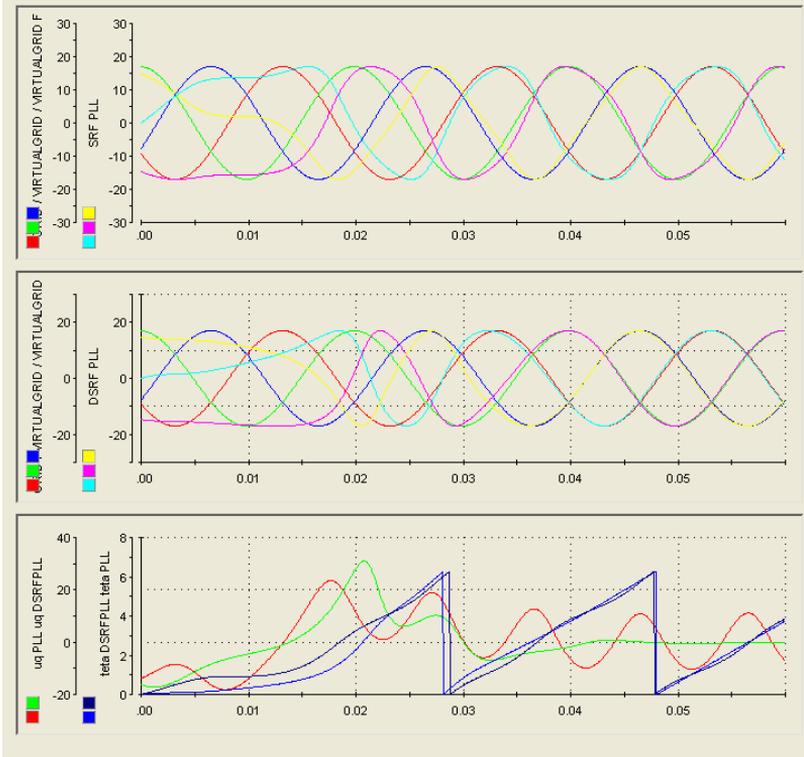
Figure 14: Synchronization in case of a synthetic unbalanced grid. The reference is the direct component of the asymmetrical voltage system. The red curve in the bottom diagram is the phase error in case of the SRF-PLL

An experiment has been performed using the output voltage system of a 3-phase 240 VA transformer, and inducing unbalance by an asymmetrical resistive load step (5A step between two phases). *Figure 15* shows the behavior of both PLL methods in this practical case. The voltage system is also distorted in this experiment, but, again, the DSRF-PLL proves to be more performant.
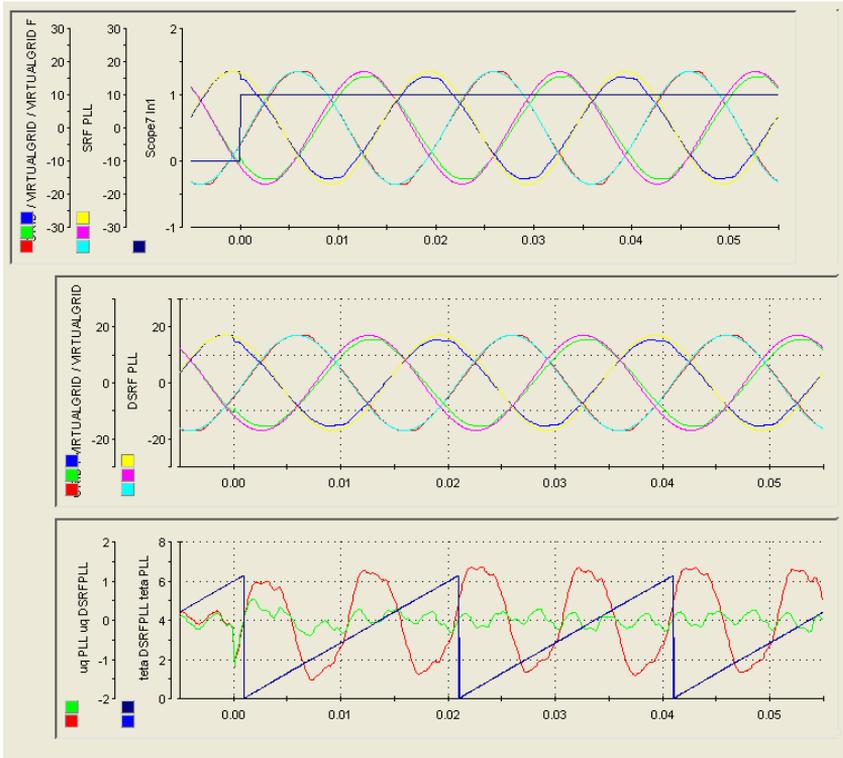
Figure 15: The response of the SRF and DSRF PLL methods to a transformer secondary voltage unbalance induced by an asymmetrical load step at t=0

5. Simulation of the switch mode rectifier with SRF and DSRF PLL synchronization, under unbalanced grid voltage conditions

This chapter compares by Matlab-Simulink simulation the PLL methods applied for the grid synchronization of a switch mode voltage rectifier shown in *Figure 10*. The discrete time models use the highest possible sampling frequency, established during the experiments for the implementation on the dS1104 real time development system.

The harmonic content of the grid currents is used to compare the performances of the PLL methods under unbalanced grid voltage conditions.

From (1) and (2) it can be observed, that the actual grid currents and voltages directly influence the synthesis of the rectifier voltage reference. On its turn, the rectifier voltages directly influence the grid currents, hence the importance of the correct identification of the direct sequence grid voltage system.

The following parameters were set:
- The converter components: $L_g = 5.6\ mH, C = 450\mu F, R_l = 32\Omega$
- DC voltage controller parameters: $K_p = 0.01, K_i = 0.1$
- Controller of the current component i_d: $K_p = 50, K_i = 10$
- Controller of the current component i_q: $K_p = 50, K_i = 10$
- Grid line voltage direct sequence component: $U_d = 380\ V_{rms}, f = 50\ Hz$
- Grid line voltage inverse sequence component: $U_i = 0.3U_d = 114\ V_{rms}, \varphi = 30°, f = 50\ Hz$
- SRF-PLL controller: $K_p = 0.584, K_i = 64.93$
- DSRF-PLL controller $K_p = 0.584, K_i = 64.93$

Figure 16 shows the unbalanced grid voltage system and the distorted grid currents in case of synchronization using the SRF-PLL method.
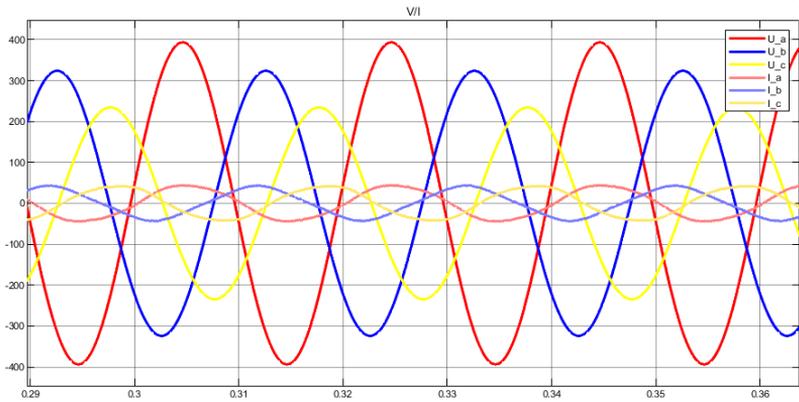
Figure 16: Asymmetrical grid and distorted currents resulting with the SRF-PLL

In *Figure 17* it can be observed that significant 3-rd and 5-th order harmonics appear in the grid current due to the fact that the SRF-PLL is not suitable for the direct sequence grid voltage component detection. In this case, the total harmonic distortion of the grid current is higher than 5%
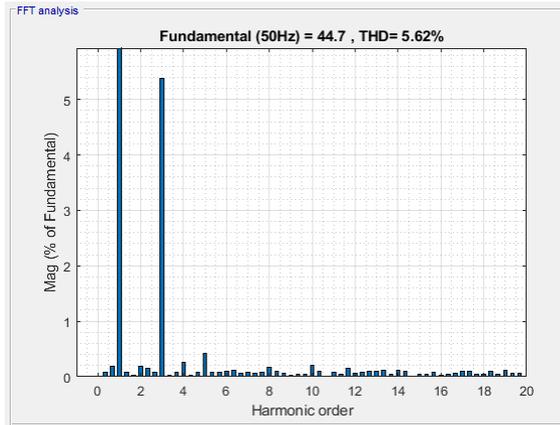
Figure 17: FFT analysis of a grid phase current in case of SRF-PLL synchronization

The result is much better when the DSRF-PLL is applied for synchronization. This is demonstrated in *Figure 18* and *Figure 19,* which show the almost sinusoidal grid currents and their harmonic contents under the same unbalanced grid voltage conditions. In this case the total harmonic distortion is close to 1.7%, and unity power factor operation can be observed as well
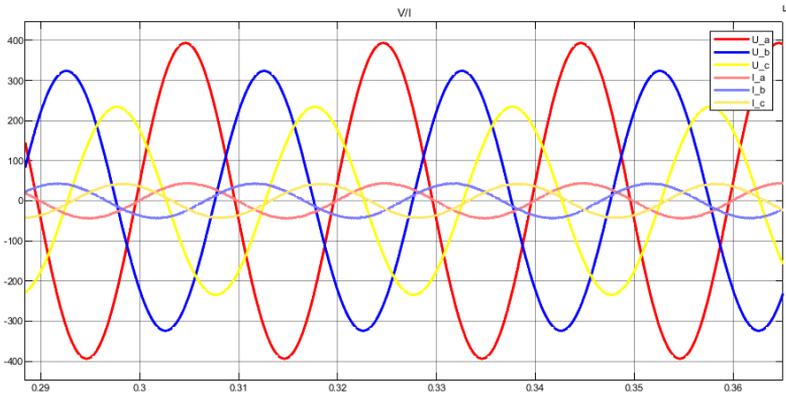
Figure 18: Asymmetrical grid voltage and symmetrical grid current system resulted when using the DSRF-PLL synchronization method
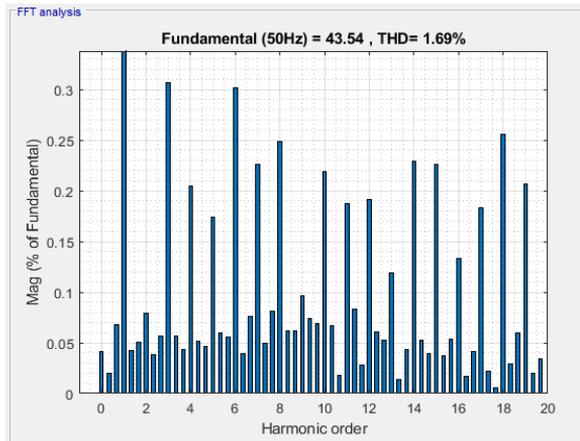
Figure 19: FFT analysis of a phase current in case of DSRF-PLL synchronization

6. Conclusions

In this study, the SRF PLL and DSRF PLL grid synchronization methods, used for the control of grid connected switch mode rectifiers, have been presented and compared. The voltage oriented control structure of the rectifier, that applies these synchronization algorithms, has been implemented on a dSpace 1104 type real time development system, and we found that a 12 kHz sampling frequency could be reached, with further improvement possibility by creating a multirating model.

A discrete time Matlab-Simulink model of the rectifier supplied from an unbalanced grid voltage system was created, which made possible the comparison of the harmonic contents of the grid currents absorbed in case of the two different synchronization methods.

For a ratio of the inverse and direct sequence grid voltage components of 0.4, we found that the total harmonic distortion of the grid currents was 5.62% in case of the application of the SRF PLL, but it decreased to 1.69 % when we applied the DSRF PLL synchronization method.

The harmonic content of the grid currents proves that in the case of an asymmetrical voltage system, the DSRF PLL provides a more performant operation of the rectifier due to its ability to accurately separate the direct and inverse sequence voltage components.

References

[1] Teodorescu, R., Liserre, M., Rodriguez, P., "Grid Synchronization in Three Phase Power Converters," in *Grid Converters for Photovoltaic and Wind Power System*s, IEEE, pp.169–204, 2007.

[2] Rodriguez, P., Pou, J., Bergas, J., Candela, I., Burgos, R., and Boroyevich, D., "Double Synchronous Reference Frame PLL for Power Converters Control", *IEEE 36th Power Electronics Specialists Conference*, 2005, pp. 1415–1421.

[3] Malinowski, M., Kazmierkowski, M. P., Trzynadlowski, A., "Review and Comparative Study of Control Techniques for Three-phase PWM Rectifiers", *Mathematics and Computers in Simulation*, 63, pp. 349–361, 2003.

[4] Kulkarni, A., and John, V., "Design of Synchronous Reference Frame Phase-Locked Loop with the Presence of DC Offsets in the Input Voltage", *IET Power Electronics*, vol. 8, no. 12, pp. 2435–2443, 2015.

[5] Rodriguez, P., Pou, J., Bergas, J., Candela, J. I., Burgos, R. P., and Boroyevich, D., "Decoupled Double Synchronous Reference Frame PLL for Power Converters Control", *IEEE Transactions on Power Electronics*, vol. 22, no. 2, pp. 584–592, 2007.

[6] Sevilmiş, F., and Karaca, H., "Performance Analysis of SRF-PLL and DDSRF-PLL Algorithms for Grid Interactive Inverters," *International Advanced Researches and Engineering Journal*, pp. 116–122, 2019.

[7] Guo, X. et al., "Phase Locked Loop and Synchronization Methods for Grid- Interfaced Converters: a Review", *Przegląd Elektrotechniczny,* pp. 182–187, 2011.

[8] dSPACE GmbH, "Real-Time Interface (RTI and RTI-MP) -Implementation Guide," Relase 4.2 –March 2005.

[9] Sanjuan, S. L., "Voltage Oriented Control of Three-Phase Boost PWM Converters", Göteborg, Sweden, Chalmers University of Technology, 2010.

[10] Rodriguez, P., Sainz, L. and Bergas, L., "Synchronous Double Reference Frame PLL Applied to a Unified Power Quality Conditioner", in Proc. *10th International Conference on Harmonics and Quality of Power*, vol. 2, 2002, pp. 614–619.

[11] Hartani, K, and Miloud, Y., "Control Strategy for Three Phase Voltage Source PWM Rectifier Based on the Space Vector Modulation", *Advances in Electrical and Computer Engineering*, 2010.

[12] Chung, S.-K., "A Phase Tracking System for Three Phase Utility Interface Inverters", *IEEE Transactions on Power Electronics*, vol. 15, no. 3, pp. 431-438, May 2000.

[13] Rodriguez, J. R., Dixon, J. W., Espinoza, J. R., Pontt, J., and Lezana, P., "PWM Regenerative Rectifiers: State of the Art", *IEEE Transactions on Industrial Electronics*, vol. 52, no. 1, pp. 5–22, Feb. 2005.

[14] Sharma, D., Sen, B., and Babu, B. C., "Improved Grid Synchronization Algorithm for DG System Using DSRF PLL Under Grid Disturbances", *2012 Students Conference on Engineering and Systems*, 2012, pp. 1–6.

[15] Akagi, H., Hirokazu, E., Aredes, M, "Instantaneous Power Theory and Applications to Power Conditioning", IEEE Press, 2007.

Theoretical Study of the Unconstrained and Constrained Nonlinear Optimal Discrete Time State Feedback Control

Katalin GYÖRGY[1], László DÁVID[2]

Sapientia Hungarian University of Transylvania, Cluj-Napoca,
Faculty of Technical and Human Sciences, Târgu Mureș,
Department of Electrical Engineering,
e-mail: [1]kgyorgy@ms.sapientia.ro, [2]ldavid@ms.sapientia.ro

Abstract: Majority of the optimal control techniques can only be applied successfully if the model of the controlled process is known and it is linear. If the system model is nonlinear, then this nonlinear model can be approximated with different simple, linear models. However, these models are valid only in the neighbourhood of the operating points. The success of the control algorithms is highly dependent on the used linearization methods. The aim of the paper is to compare different optimal control algorithms and linearization methods. The presented optimal control algorithms have been also tested in constrained and unconstrained versions.

Keywords: LQR control, SDRE control, nonlinear system, cost function, constrained control, parameter estimation

1. Introduction

The nonlinear optimal control, namely constrained optimal control of nonlinear dynamical systems, still remains nowadays a very interesting, useful and constantly evolving field of science. Its origin has been laid down in the 1950's with the introduction of dynamic programming, the Hamilton-Jacobi-Bellman partial differential equations, and the Pontryagin maximum-minimum principle. From these beginnings, numerous design methodologies have been developed, from the direct solution through numerical computations, generalizations of the classical Lyapunov theory with the control Lyapunov function (CLF), and extending the linear optimal control theory with linearization and the state dependent Riccati equation (SDRE) based techniques.

This paper briefly discusses the possible applications of nonlinear optimal controllers. There are presented some control algorithms that require to know the state space mathematical model of the controlled process, and an adequate criteria function. The known optimal control theories can be applied successfully to control of processes characterized by linear system models. The question arises, how can these approaches be applied to processes characterized by nonlinear system models? There exists a well-known method of linearization around the operation point, which requires the definition of the Jacobian matrix, which in many cases is disadvantageous. In this paper two linearization methods are presented, one of which requires knowledge of the analytical model of the system, and the other does not. This second version considers a black box model of the process to be controlled. It is important that all state variables have to be accessible. In both cases, the result will be a so-called linearized mathematical model, where the specified model matrices are state dependent. Two types of controllers will be introduced, where the optimal criteria function is quadratic: Discrete Linear Quadratic Regulator (DLQR) and the advanced Model Predictive Control (MPC) algorithm, where some constraints can also be introduced [1].

The present paper has the following structure. Section 2 briefly discusses different linearization methods, and section 3 introduces two control algorithms. Section 4 demonstrates the applicability of the presented methods using the nonlinear dynamical model of the inverted pendulum. The short conclusions are outlined in section 5.

2. Nonlinear optimal control theory

In this section, we present the nonlinear discrete time control, and its corresponding implementation theory. Consider the following discrete nonlinear mathematical model [2], [3]

$$
\begin{aligned}
x_{k+1} &= F(x_k, u_k) \\
y_k &= G(x_k, u_k)
\end{aligned}
\tag{1}
$$

where x_k is the n dimensional state vector, u_k is the m dimensional input vector and y_k is the p dimensional output vector. The functions $F{:}R^{n+m}{\rightarrow}R^n$ and $G{:}R^{n+m}{\rightarrow}R^p$ are continuous, nonlinear vector functions. Thereinafter the state feedback control theory will be presented. The investigation of a tracking control problem is similar, but in the tracking control case, there appear some extra terms. Since this paper is only about the state feedback control, in the following the output equation will not be considered.

For linearization, this mathematical model (1) has to be transformed to the following mathematical form:

$$x_{k+1} = A(x_k) \cdot x_k + B(x_k) \cdot u_k \tag{2}$$

where the state-space system matrices (A and B) are state-dependent.

A. *Linearization around the operating point*

The nonlinear optimal control methods usually involve Jacobian linearization of the system model around each operation point. This method assumes that, the discrete mathematical model and the parameters are known and the nonlinear functions (in the model description) are continuous. The linearized matrices can be computed by the following relations.

$$A(x_k) \cong \begin{bmatrix} \dfrac{\partial F_1}{\partial x_1} & \cdots & \dfrac{\partial F_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial F_n}{\partial x_1} & \cdots & \dfrac{\partial F_n}{\partial x_n} \end{bmatrix}_{x_k,u_k} \quad B(x_k) \cong \begin{bmatrix} \dfrac{\partial F_1}{\partial u_1} & \cdots & \dfrac{\partial F_1}{\partial u_m} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial F_n}{\partial u_1} & \cdots & \dfrac{\partial F_n}{\partial u_m} \end{bmatrix}_{x_k,u_k} \tag{3}$$

This linear approximation of the model may introduce errors, especially if the states are far from the equilibrium point. In addition, the numerical approach of the derivatives can also bring inaccuracies [2].

B. *Discrete time SDRE method*

According to the SDRE theory, this can be easily accomplished if the F function in (1) can be written as [8], [13].

$$x_{k+1} = F(x_k, u_k) = f(x_k) + g(x_k) \cdot u_k \tag{4}$$

by introducing the following formal notation:

$$A(x_k) = \frac{f(x_k)}{x_k}, \quad B(x_k) = g(x_k) \tag{5}$$

This method can be used if the plant is characterized by a known nonlinear mathematical model with known parameters and this mathematical model can be transformed in a special difference equation form (4).

C. *Linearized model with parameter estimation*

This solution does not require knowledge of the mathematical model of the system, but states must be measurable here as well. In this case we determine the linearized matrices based just on the measured values. For this method the model (2) will be transformed theoretically as follows:

$$x_k^T = \begin{bmatrix} x_{k-1}^T & u_{k-1}^T \end{bmatrix} \cdot \begin{bmatrix} A^T(x_{k-1}) \\ B^T(x_{k-1}) \end{bmatrix} = \varphi_k^T \cdot \theta_k \tag{6}$$

The well-known parametric estimation algorithm can be applied to this mathematical model form, where θ_k contains the unknown parameters and φ_k is the measurement vector. The following steps characterize the Least Square Estimation (LSE) recursive algorithm [10]:

$$\begin{aligned} \varepsilon_k &= x_k^T - \varphi_k^T \cdot \theta_k, \\ K &= \frac{F_k \cdot \varphi_k}{\lambda + \varphi_k^T \cdot F_k \cdot \varphi_k}, \\ F_{k+1} &= \frac{1}{\lambda} \cdot (F_k - K \cdot \varphi_k^T \cdot F_k), \\ \theta_{k+1} &= \theta_k + K \cdot \varepsilon_k, \end{aligned} \tag{7}$$

where ε_k is the estimation error, K is the estimation gain vector, F_k the covariance matrix, λ the forgetting factor and θ_k is the estimated parameter vector. This is a recursive algorithm, so we have to make some initialization first. There have to be chosen the initial value of the parameter vector θ_0 and the initial value of the covariance matrix F_0.

3. Control methods

This section briefly introduces two control approaches. The first method is the DLQR method, which is not a strictly constrained control algorithm, with this only the weak constraints can be specified. Here the idealized infinite horizon control and the finite horizon version can also be tested. The second method is the MPC algorithm, which takes into account also the strict constraints, when calculating the control signal.

A. DLQR

Here only the state feedback control is discussed, where the goal is to control all states to zero. This type of problem is characterized with the following discrete criteria function [2], [3]:

$$J(\Delta u_k) = \sum_{k=0}^{N} x_k^T \cdot Q_k \cdot x_k + \sum_{k=0}^{N-1} \Delta u_k^T \cdot R_k \cdot \Delta u_k \tag{8}$$

where N is the horizon value and here it is considered, that the weight matrices R and Q are not state dependent. The variation of control is defined as

$$\Delta u_k = u_k - u_{k-1}. \tag{9}$$

In a similar way to the classical LQR design the variation of the state-feedback control is calculated as:

$$\Delta u_k = -\left(R + B_k^T \cdot P_k \cdot B_k\right)^{-1} \cdot \left(B_k^T \cdot P_k \cdot A_k \cdot x_k + B_k^T \cdot P_k \cdot u_{k-1}\right) \tag{10}$$

For simplicity the following notations have been introduced for the Riccati matrix and for the state dependent matrices: $P(x_k)=P_k$, $A(x_k)=A_k$ and $B(x_k)=B_k$.. The Riccati matrix is the unique solution of the following discrete time state dependent Riccati equation.

$$P_k = Q + A_k^T \cdot (P_{k+1} - P_{k+1} \cdot B_k \cdot \left(R + B_k^T \cdot P_{k+1} \cdot B_k\right)^{-1} B_k^T \cdot P_{k+1}) \cdot A_k \tag{11}$$

The solution of SDRE (11) is only a sub-optimal solution, because there were neglected derivatives of the system matrices ($A(x_k)$ and $B(x_k)$) [8]. If the value of the horizon (N) is infinite, the solution of the difference Riccati equation approaches the solution of the algebraic equation ($P_k = P_{k+1}$) [2].

B. MPC

The model predictive algorithm looks for the vector ΔU_k that minimizes a cost function represented by the following relationship

$$J(\Delta U_k) = X_k^T \cdot Q^* \cdot X_k + \Delta U_k^T \cdot R^* \cdot \Delta U_k \tag{12}$$

where X_k is the vector with the predictions of the controlled state variables, ΔUk is a vector with future input changes, $Q*$ is a diagonal matrix with weights for the states, $R*$ is a diagonal matrix with weights for the control action changes [1], [7], [15]. If the prediction horizon is N and the control horizon is Nc, these vectors and matrices are:

$$X_k = \begin{bmatrix} x_{k+1/k} \\ \vdots \\ x_{k+N/k} \end{bmatrix} \quad \Delta U_k = \begin{bmatrix} \Delta u_{k/k} \\ \vdots \\ \Delta u_{k+N_c-1/k} \end{bmatrix} \tag{13}$$

$$Q^* = \begin{bmatrix} Q_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & Q_N \end{bmatrix} \quad R^* = \begin{bmatrix} R_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & R_{N_c-1} \end{bmatrix} \tag{14}$$

The following representation is obtained for the predictions:

$$X_k = A_k^* \cdot x_k + B_k^* \cdot u_{k-1} + G_k^* \cdot \Delta U_k \tag{15}$$

where

$$A_k^* = \begin{bmatrix} A_k \\ \vdots \\ A_k^N \end{bmatrix} \quad B_k^* = \begin{bmatrix} B_k \\ \vdots \\ \sum_{i=0}^{N-1} A_k^i \cdot B_k \end{bmatrix}, G_k^* = \begin{bmatrix} B_k & \dots & 0 \\ \vdots & \ddots & \vdots \\ \sum_{i=0}^{N-1} A_k^i \cdot B_k & \dots & \sum_{i=0}^{N-Nc} A_k^i \cdot B_k \end{bmatrix} \qquad (16)$$

The cost function can be written as

$$J(\Delta U_k) = \frac{1}{2} \Delta U_k^T \cdot H \cdot \Delta U_k + f^T \cdot \Delta U_k + const \qquad (17)$$

where

$$H = 2 \cdot \left[G_k^T \cdot Q^* \cdot G_k + R^* \right] \qquad (18)$$

$$f = 2 \cdot G_k^T \cdot Q^* \cdot (A_k^* \cdot x_k + B_k^* \cdot u_{k-1}) \qquad (19)$$

For the unconstrained problems, the model predictive control determines the vector ΔU_k:

$$\Delta U_k = \frac{1}{2} \cdot \left(H + H^T \right)^{-1} \cdot f =$$
$$= -\left(G_k^T \cdot Q^* \cdot G_k + R^* \right)^{-1} \cdot G_k^T \cdot Q^* \cdot (A_k^* \cdot x_k + B_k^* \cdot u_{k-1}) \qquad (20)$$

To the quadratic cost function (17) there can be assigned different linear constraint inequalities

$$A_c \cdot \Delta U_k \le B_c \qquad (21)$$

These types of problems can be solved with different numerical quadratic programming algorithms (ex. interior point method, quadratic penalty method [4], [5], [6]).

4. Example and numerical simulation

For testing the presented control methods, the well-known cart on inverted pendulum dynamical system can be used as an example. The control of this system is quite difficult due to the characteristics of the system: instability, nonlinearity of the model, with single input and four state variables. The schematic representation of the inverted pendulum system is shown in *Fig. 1*.
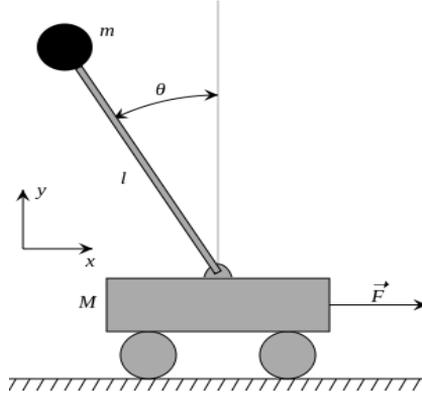
Figure 1: Inverted pendulum on cart

The Euler-Lagrange theory has been used to determine this mathematical model, where the inertia of the system has been neglected [9], [11], [12]. The nonlinear mathematical discrete time model is the following.

$$x_{1,k+1} = x_{1,k} + T_s \cdot x_{2,k}$$

$$x_{2,k+1} = x_{2,k} + T_s \cdot \frac{\left(l \cdot x_{4,k}^2 \cdot \sin(x_{3,k}) - g \cdot \sin(x_{3,k}) \cdot \cos(x_{3,k}) + \frac{1}{m} \cdot u_k \right)}{(\frac{M}{m} + \sin^2(x_{3,k}))}$$

$$x_{3,k+1} = x_{3,k} + T_s \cdot x_{4,k} \tag{22}$$

$$x_{4,k+1} = x_{4,k} + T_s \cdot \frac{\left(\frac{m+M}{ml} g \cdot \sin(x_{3,k}) - \frac{1}{2} x_{4,k}^2 \cdot \sin(2 \cdot x_{3,k}) - \frac{1}{ml} \cdot \cos(x_{3,k}) \cdot u_k \right)}{(\frac{M}{m} + \sin^2(x_{3,k}))}$$

The four state variables in the model are: $x_{1,k}$ displacement of the cart, $x_{2,k}$ speed of the car, $x_{3,k}$ pendulum angle and $x_{4,k}$ pendulum angular speed. The input signal u_k is the force acting horizontally on the car. The model parameters and their numeric values are shown in Table 1.

Table 1: Model parameters

System parameters	Value
M (mass of cart)	0.6 kg
m (mass of pendulum)	0.45 kg
l (length of pendulum)	0.35 m
g (acceleration of gravity)	9.81 m/s^2
Ts (time sampling)	0.01 s

In the following there will be listed some special values and relationships that will be used along with different linearization methods:

- For the linearization around the operating point it is necessary to calculate certain derivatives. This can be solved using a simple approximation method, where the perturbation values is set to 0.0001.

- For testing the SDRE method the following state dependent matrices can be used:

$$A(x_k) = \begin{bmatrix} 1 & T_s & 0 & 0 \\ 0 & 1 & T_s \cdot \dfrac{-g \cdot \sin(x_{3,k}) \cdot \cos(x_{3,k})}{x_{3,k} \cdot (\frac{M}{m} + \sin^2(x_{3,k}))} & T_s \cdot \dfrac{l \cdot x_{4,k} \cdot \sin(x_{3,k})}{(\frac{M}{m} + \sin^2(x_{3,k}))} \\ 0 & 0 & 1 & T_s \\ 0 & 0 & T_s \cdot \dfrac{\frac{m+M}{ml} g \cdot \sin(x_{3,k})}{x_{3,k} \cdot (\frac{M}{m} + \sin^2(x_{3,k}))} & 1 - T_s \cdot \dfrac{\frac{1}{2} x_{4,k} \cdot \sin(2 \cdot x_{3,k})}{(\frac{M}{m} + \sin^2(x_{3,k}))} \end{bmatrix} \quad (23)$$

$$B(x_k) = \begin{bmatrix} 0 \\ \dfrac{\frac{1}{m} \cdot T_s}{(\frac{M}{m} + \sin^2(x_{3,k}))} \\ 0 \\ \dfrac{-T_s \cdot \frac{1}{ml} \cdot \cos(x_{3,k})}{(\frac{M}{m} + \sin^2(x_{3,k}))} \end{bmatrix} \quad (24)$$

- For the parameter estimation algorithm, the choice of the initial parameter matrix and of the initial covariance matrix is important. For better results, the initial value of the parameter matrix is initialized with the initial values obtained by the SDRE method and the initial covariance matrix is a diagonal matrix $10 \cdot I_5$.

For the DLQR and the MPC control methods, the following parameters and limits have been used: the weighting matrix (value) of the controls $R=0.1\ I_1$, the weighting matrix of the states $Q=100\ I_4$, the prediction horizon $N=50$, the control horizon $Nc=5$, and the limits of the control signals $u_{min}=-50$ and $u_{max}=50$. Here can be set also some limits for the variation of the control $\Delta u_{min}=-10$ and $\Delta u_{max}=10$. In accordance with (21) these limits can be given by the following matrices:

$$A_c = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ -1 & 0 & \cdots & 0 \end{bmatrix} \quad B_c = \begin{bmatrix} u_{\max} - u_{k-1} \\ -u_{\min} + u_{k-1} \\ \Delta u_{\max} \\ -\Delta u_{\min} \end{bmatrix} \quad (25)$$

For numerical simulation the following abbreviations were used to refer to the presented linearization methods:

L_op – method of linearization around an operating point;

L_sdre - linearization using the SDRE approach;

L_est - linear parameter estimation of the discrete time linear model.

The following notations will be assigned to the type of control algorithm: DLQR1 - optimal quadratic control algorithm with infinite horizon, DLQR2 - optimal quadratic control algorithm with finite horizon and MPC- model predictive control.

The graphical results of the unrestricted control version (for the presented control methods and linearization methods) are presented in the following figures. *Fig. 2* shows the control result (variation of the states $x_1(t)$ and $x_3(t)$) achieved by the linearization method around the operation point. These figures show the control effects of the initial states and of the modified states during the simulation (at about 40 sec). In this linearization method (L_op) the weakest control result was obtained for the case of finite horizon method. The best result was obtained for the infinite horizon DLQR control, but this method can be used just for theoretical simulations. The MPC algorithm is also a good solution, but here is already used the control horizon value (N_c). All this is observed mainly in the controlling of the initial state. The perturbation of states (during the simulation) leads to similar results in all three variants.
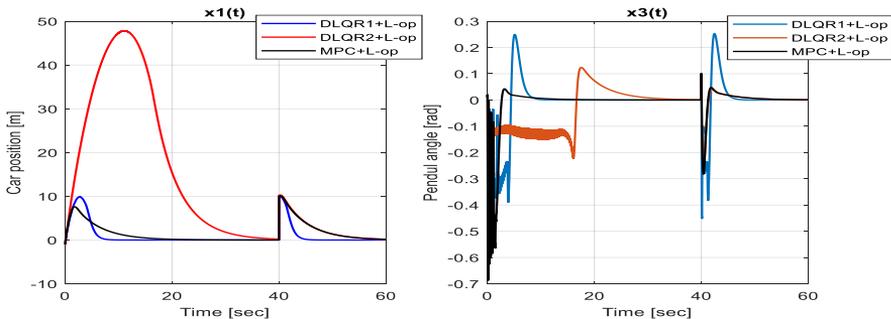
Figure 2: States controlled by different unconstrained control methods using linearization around operation points (L_op)
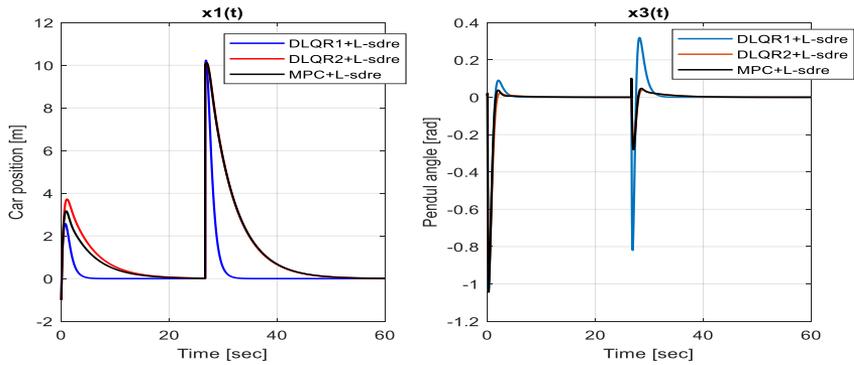
Figure 3: States controlled by different unconstrained control methods using linearization based on SDRE approach (L_sdre)
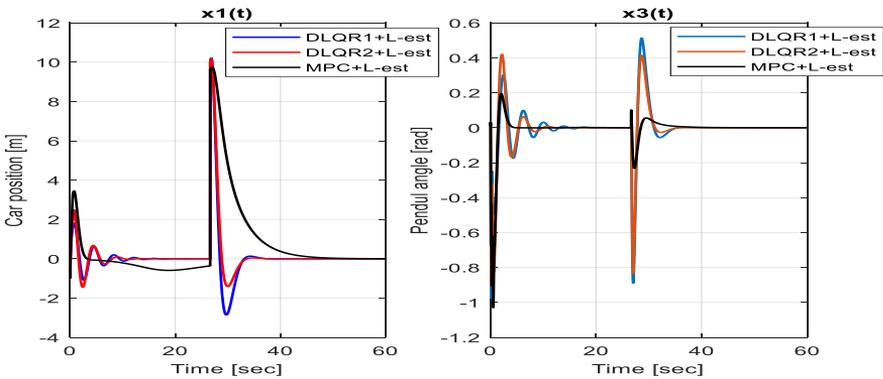
Figure 4: States controlled by the different unconstrained control methods using the linearization method with parameter estimation (L_est)

Fig. 3 and *Fig. 4* in principle the same process is shown, but here the linearization is made with SDRE approach and with parameter estimation algorithms. These plots show also the effect on the control process both of the initial state and of the perturbation at about 30 sec. Here the perturbation of states leads to similar results in all three variants. These simulation results have been obtained for unconstrained control signals. The variation of the control signals (for the first 30 sec) in case of using the third linearization method (L_est) is shown in *Fig. 5*.
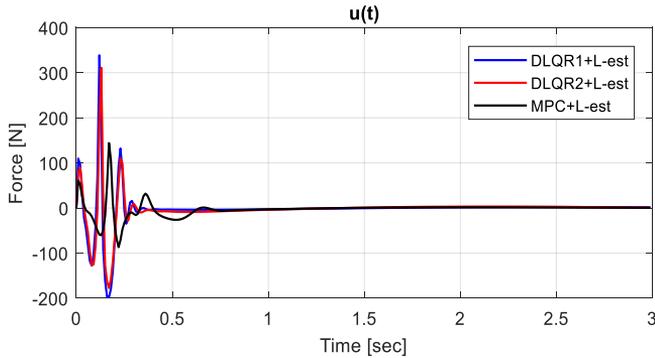
Figure 5: Variation of the unconstrained control signals using the linearization method with parameter estimation (L_est)

In the following we test the effect of the constrained control on these algorithms. Taking into account the limits of the control signal (and their variations), the results are shown in *Fig. 6* and *Fig. 7*. For the MPC algorithm, these constraints are already taken into account in the calculation of the control signal. In the other methods (LQR), the signal values are simply limited when the thresholds are exceeded. The constrained control results for the LQR1 algorithm are shown in *Fig. 6* and for MPC algorithm in *Fig. 7*. In both cases, the linearized model is determined by an on-line parametric estimation method. It can be seen from these figures that the MPC algorithm handles much better the constrained control problems.
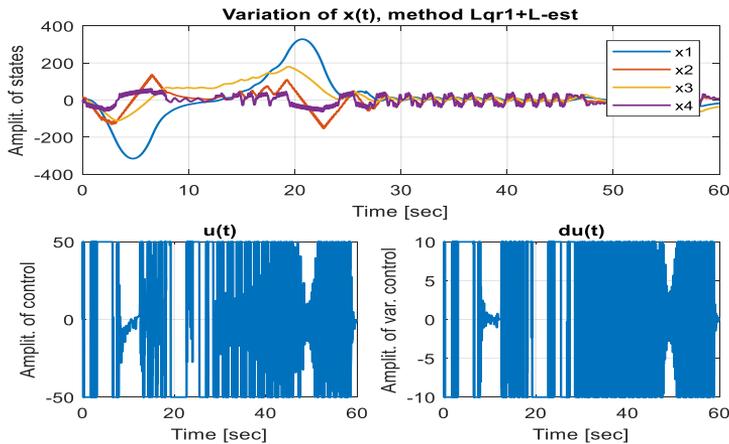
Figure 6: States controlled by constrained LQR method, where the linearized parameter estimation (L_est) was used
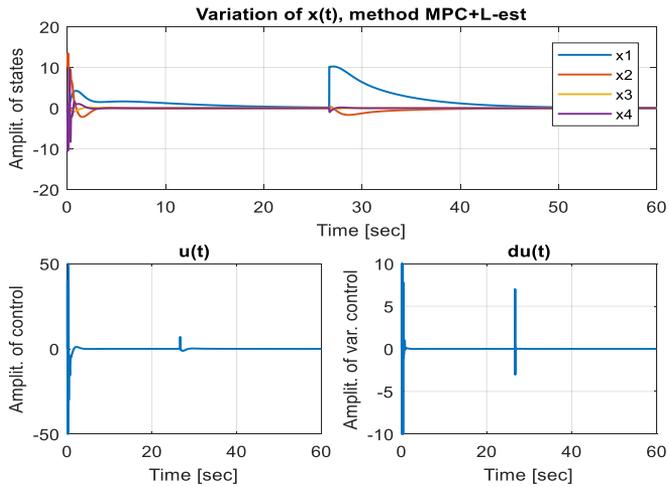
Figure 7: The evolution in time of the states controlled with constrained MPC
method, with the linearized parameter estimation (L_est) method

The presented algorithms have been tested also for small changes of the system parameters. The first two linearization methods (L_op and L_sdre) do not have a good behavior in case of the parametric change. This is because these two methods require the knowledge of the analytical model of the process. However, the third linearization method (L_est) has a proper effect in eliminating the unpleasant effect of such a modification (adaptive control behavior). This is shown in *Fig. 8* where during the simulation (after about 50 sec) some of the system parameters have been modified. Here it is shown the variation of the controlled states, and the change of the estimated during the estimation algorithm can also be observed.
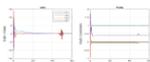
Figure 8: Controlled states and estimated parameters
(MPC method with linearized parameter estimation)

5. Conclusion

In the paper a theoretical study has been presented regarding the implementation off different state feedback optimal control approaches in case of a nonlinear system model. Two types of optimal controller have been presented: the LQR controller and the modern MPC controller. Because the goal is to apply these controllers to a nonlinear system, there is need for linearization. To accomplish this, there have been presented three possible linearization methods: linearization around the operation point, linearization based on the SDRE approach, and finally a method using parameter estimation algorithm. These control algorithms and linearization methods have been tested on a nonlinear system, namely inverse pendulum. Applying these methods, it can be observed that the linearization around the operation point is very sensitive, and the convergence of the parametric algorithm depends on the correct choice of the initial values. The control algorithms are highly dependent on the control parameters, the value of the horizon and the predetermined limits. For the constrained control tasks, the MPC controller provides very good results (and it is valid, independently of the used linearization methods).

References

[1]   Camacho, E. F., "Model Predictive Control", Ed. Springer, 2004.
[2]   David, L., "Tehnici de optimizare", Petru Maior University, Tg. Mureș, 2000.
[3]   Kirk, D. E., "Optimal Control Theory: An Introduction", Englewood Cliffs, NJ: Prentice Hall, 1970, reprinted by Dover, Mineola, NY, 2004.
[4]   Goodwin, G. C., Seron, M. M. and De Dona J.A., "Costrained Control and Estimation", Springer, 2004.
[5]   Boyd, S. and Vanderberghe, L., "Convex Optimization", Cambridge Univ. Press, 2009.
[6]   Nocedal, J. and Wright, S. J., "Numerical Optimization", Springer 1999.
[7]   György, K., Dávid, L., "Comparision between unconstrained MPC and discrete time LQR algorithms with finite horizon", *XIX International Confeence on Computer Science, 2009, Tg. Mureş, 8-11 oct. 2009,* pp. 220−226.
[8]   György, K., Dávid L. and Kelemen, A., "Theoretical Study of the Nonlinear Control Algorithms with Continuous and Discrete-Time State Dependent Riccati Equation", *Inter-Eng 2015 "Petru Maior" University, Tg. Mureş, oct., 2015, Elsevier, Procedia Technology 22 (2016),* pp. 582−591.
[9]   Gyorgy, K., "The LQG Control Algorithms for Nonlinear Dynamic Systems", *Inter-Eng 2018 "Petru Maior" University, Tg. Mureş, oct, 2018, Science Direct, Procedia Technology 32 (2019),* pp. 553−563.
[10] Ljung, L., "System identification, Theory for the User", Prentice Hall, 1999.
[11] Ata, B., Coban, R., "Linear Quadratic Optimal Control of an Inverted Pendulum on a Cart using Artificial Bee Colony Algorithm. An Experimental Study", *Cukurova University Journal of the Faculty of Engineering and Arhitecture (June, 2017), 32 (2)*, pp. 109−123.
[12] Priyadarshi, P., "Optimal Controller Design for Inverted Pendulum", *Department of Electrical Engineering, National Institute of Technology, Rourkela, India, 2013, http://ethesis.nitrkl.ac.in//1/211EE3342.pdf.*
[13] Yucelen, T., Sadahalli, A. S. and Pourboghrat, F., "Online Solution of State Dependent Riccati Equation for nonlinear System Satbilization", *Amer. Control Conf., Baltimore, MD, June 2010,* pp. 6336−6341.

[14] Dutka, A. S., Ordys, A. W., Grimble, M. J. "Optimized Discrete-Time State Dependent Riccati Equation Regulator", *in Amer. Control Conf., Portland, OR, June 2005,* pp. 2293–2298.

[15] Chang, I., Bentsman, J., "Constrained Discrete-Time State-Dependent Riccati Equation Technique: A Model Predictive Control Approach", *52nd IEEE Conference on Decision and Control Dec. 10-13, 2013. Florence, Italy*, pp. 5125–5130.

Particle Swarm Optimization of a Hybrid Energy Storage System

János FERENCZ[1], András KELEMEN[2]

[1] Technical University of Cluj-Napoca, Faculty of Electrical Engineering,
Department of Electrical Machines and Drives, e-mail: janos_ferencz@yahoo.com
[2] Sapientia-Hungarian University of Transylvania, Cluj-Napoca,
Faculty of Technical and Human Sciences, Târgu Mureş,
Department of Electrical Engineering, e-mail: kandras@ms.sapientia.ro

Abstract: The paper presents the energy loss minimization of a hybrid energy storage system used in an electric vehicle, composed by a battery and a supercapacitor. The optimization is carried out by searching the optimal power sharing between the energy storage devices. The power sharing factor is defined as a discrete time variable, with constant values during each subdivision of the driving cycle. The elements of the optimal solution vector are the power sharing factors and the time instants that define the subdivisions. The particle swarm optimization algorithms have been validated using the Rastrigin test function, and three versions of the boundary behaviour have been compared in case of the constrained optimization. The algorithms have been tested for the energy loss minimization in case of a simple driving cycle, and their performance has been assessed by statistical analysis for different swarm sizes.

Keywords: Particle swarm optimization, hybrid energy storage system, electric vehicle, constrained optimization.

1. Introduction

For the efficient operation of the electric vehicles, it is crucial to combine different types of energy storage devices, due to the fact each type has a limited range of operating conditions it can efficiently handle [4]. The high energy storage capability of the batteries, necessary to guarantee a high autonomy of the vehicle, is not associated with high output power capability, which is required during acceleration and deceleration. Thus, the hybridization of the energy storage system is the solution to satisfy the electrical energy supply or storage demands in different operating states [7].

The Hybrid Energy Storage System (HESS) being studied in this paper, illustrated in *Fig. 1,* consists of the combination of a Li-ion battery and a supercapacitor [14], in an active parallel topology that allows individual or joint power delivery to the load (to the wheel motors), energy recovery in case of braking, and energy exchange between these storage devices by means of power electronic converters. The battery life cycle is extended if it's protected by the supercapacitor from current shocks during the vehicle power peaks [9], [13].
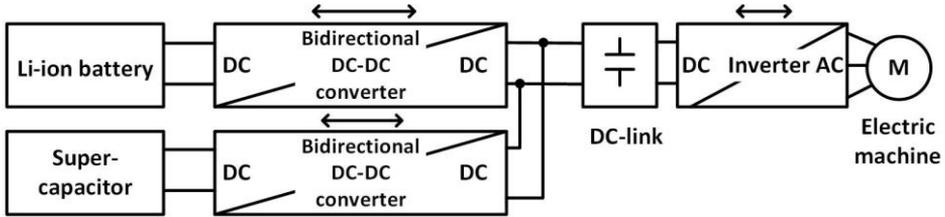
Figure 1: Block diagram of the active parallel hybrid energy storage system (HESS) [13]

The energy management of such a hybrid energy storage system has the task to determine in every instant of operation the "proper contribution" of each storage device to the power flow in either direction. This term can be converted to an objective function to be optimized under different constraints. One can aim to maximize the life cycles of the storage devices, to maximize the vehicle's autonomy, to minimize the operation costs, etc. Limitations can possibly be imposed by state of charge or state of energy minima, or by maximum admissible currents, but vehicle-to-grid system requirements regarding the state of charge at peak or off-peak hours might also be feasible constraint.

Fig. 2 shows the structure of the model used to study the energy management algorithm (EMA), which provides p_{BAT} and p_{SC}, i.e. the powers to be delivered by the battery and the supercapacitor, respectively, based on the following inputs:
- battery state of charge (SOC);
- instantaneous battery current, voltage, and internal resistance $(i_{BAT}, u_{BAT}, r_{BAT})$;
- supercapacitor state of energy (SOE);
- instantaneous supercapacitor current, voltage, and internal resistance (i_{SC}, u_{SC}, r_{SC});
- instantaneous electrical power required to drive the vehicle (p_{el_req}).

The aim of this paper is the minimization of the energy losses on the internal resistances of the battery and of the supercapacitor. The higher efficiency contributes also the increase of the vehicle's autonomy.
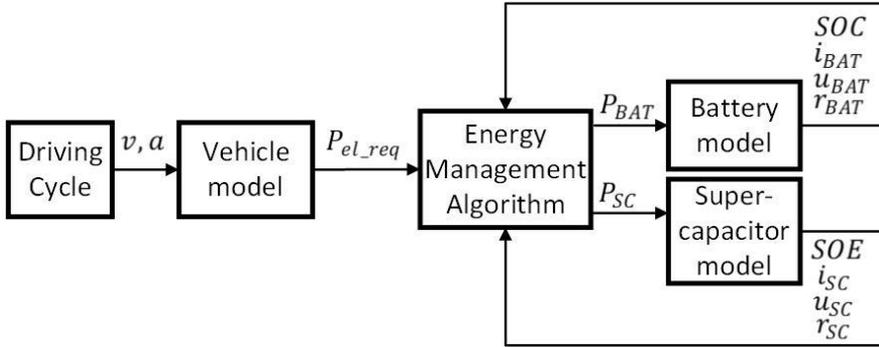
Figure 2: The block diagram of the model including the energy management algorithm of the hybrid energy storage system

The task of the EMA is to determine the optimal power sharing between the battery and the supercapacitor, considering their power limitations, and thus extending their life cycles.

2. The optimization problem

The performance of the vehicle is defined for standard driving cycles, representative for different characteristic operating conditions, usually formulated as theoretically or practically derived time-diagrams of speed. Thus, also the acceleration can be determined in each moment to derive the mechanical power required by the given speed profile. As we don't aim to detail the mechanical model of the vehicle, let's consider that the required mechanical power at a given moment is a function of the instantaneous speed v_v and acceleration a_v, and the required electrical power p_{el_req} results from this by considering a global efficiency η_v [10], according to (1):

$$p_{el\_req}(t) = \frac{1}{\eta_v} f\big(v_v(t), a_v(t)\big). \tag{1}$$

The question is how to share this power between the energy storage devices in a way that ensures maximum efficiency, i.e. minimum power losses over a driving cycle. Generally, there should exist an optimum function of time of the sharing, which corresponds to this goal. Let's denote by x the share of the supercapacitor (either delivered or absorbed), defined by (2).

According to this definition, the power has the same sign at the terminals of both devices, and the case of energy exchange between them is not covered in this paper.

$$x(t) = \frac{p_{SC\_req}(t)}{p_{SC\_req}(t) + p_{BAT\_req}(t)}, \qquad x \in [0,1]. \tag{2}$$

Thus, the time dependent powers required from the supercapacitor and the battery are

$$\begin{cases} p_{SC\_req} = x p_{el\_req} \\ p_{BAT\_req} = (1 - x) p_{el\_req}, \end{cases} \tag{3}$$

The powers delivered by the supercapacitor and the battery result by taking into account the internal losses:

$$\begin{cases} p_{BAT} = u_{BAT} i_{BAT} - r_{BAT} i_{BAT}^2 \\ p_{SC} = u_{SC} i_{SC} - r_{SC} i_{SC}^2, \end{cases} \tag{4}$$

where in the model u_{BAT} and r_{SC} are considered constants. The supercapacitor voltage u_{SC} is time-dependent according to (5), wich uses the notation from *Fig. 3*:

$$u_{SC}(t) = u_{SC}(0) - \frac{1}{C_{SC}} \int_0^t i_{SC} dt \tag{5}$$
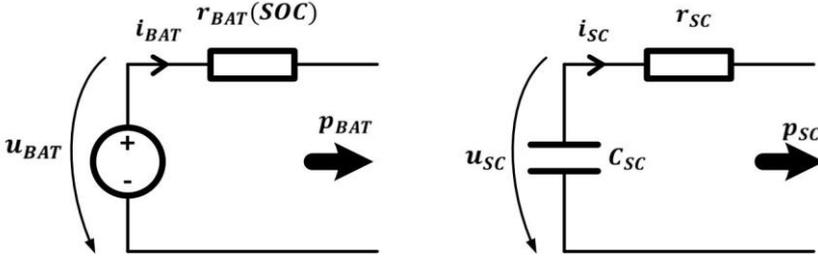
Figure 3: The equivalent circuits of the energy storage devices

We considered that the battery depletion is limited to $SOC = 50\%$ (i. e. to half of the maximum stored charge Q_{max}), thus the linear increase of its internal resistance with the decrease of the SOC, according to (6), is a viable assumption [1], [8], [11].

$$r_{BAT}(SOC) = r_{BAT}|_{SOC=50\%} - (r_{BAT}|_{SOC=50\%} - r_{BAT}|_{SOC=100\%})(SOC - 50)/50 \tag{6}$$

Thus, (3) - (6) yield the currents of the storage devices:

$$\begin{cases} i_{BAT} = \dfrac{u_{BAT} - \sqrt{u_{BAT}^2 - 4r_{BAT}(1-x)p_{el_{req}}}}{2r_{BAT}} \\ i_{SC} = \dfrac{u_{SC} - \sqrt{u_{SC}^2 - 4r_{SC}xp_{el\_req}}}{2r_{SC}} \end{cases}. \tag{7}$$

The total power loss at a given instant is

$$p_{loss} = \frac{\left(u_{BAT} - \sqrt{u_{BAT}^2 - 4r_{BAT}(1-x)p_{el_{req}}}\right)^2}{4r_{BAT}} + \frac{\left(u_{SC} - \sqrt{u_{SC}^2 - 4r_{SC}xp_{el\_req}}\right)^2}{4r_{SC}}, \tag{8}$$

while the energy losses until that moment are

$$W_{loss}(t) = \int_0^t p_{loss} dt. \tag{9}$$

A possible choice for the objective function to be minimized can be the instantaneous power loss (10), which enables online optimization, according to the procedure from *Fig. 4*.

$$J_p = p_{loss}(t) \tag{10}$$

As intuitively expected, and further on demonstrated in the paper, better results can be obtained according to the procedure from *Fig. 5* when the objective function is the total energy loss at $t = T$ (11), i.e. at the end of the driving cycle.

$$J_W = W_{loss}(T) \tag{11}$$

According to the procedure from *Fig. 4*, the instantaneous power loss (i.e. the objective function in this case) is calculated numerically using the discretization step T_s. The power loss minimization is performed on the one dimensional solution space of the instantaneous scalar power sharing factor x. The calculation of the energy loss is made by the integration of the instantaneous power minima to the end of the driving cycle.

The optimization from *Fig. 5* is made on the multidimensional solution space of the power sharing factors that correspond to each time step of the driving cycle. The dimensionality of this problem is not manageable, and the number of components of the solution vector must be reduced as discussed in the further sections.

$$\text{Init: } [W_{loss}, t, u_{SC0}, SOC, i_{BAT}, i_{SC}] \mid t = 0$$

$$t^{<k>} = t^{<k-1>} + T_s; \text{ stop if } t^{<k>} \geq T$$

$$v^{<k>}, a^{<k>}$$

$$p_{mech}^{<k>}$$

$$p_{el\_req}^{<k>} \qquad (1)$$

$$p_{BAT}^{<k>}(x^{<k>}), p_{SC}^{<k>}(x^{<k>}) \qquad (3)$$

$$SOC^{<k>} = SOC^{<k-1>} - i_{BAT}^{<k-1>} \frac{T_s}{Q_{max}}$$

$$u_{SC}^{<k>} = u_{SC}^{<k-1>} - i_{SC}^{<k-1>} \frac{T_s}{C_{SC}}$$

$$r_{BAT}^{<k>}(SOC^{<k>}) \qquad (6)$$

$$i_{SC}^{<k>}(u_{SC}^{<k>}, p_{SC}^{<k>}) \qquad (7)$$

$$i_{BAT}^{<k>}(r_{BAT}^{<k>}, p_{BAT}^{<k>}) \qquad (7)$$

$$p_{loss\_SC}^{<k>}(i_{SC}^{<k>})$$

$$p_{loss\_BAT}^{<k>}(r_{BAT}^{<k>}, i_{BAT}^{<k>})$$

$$p_{loss}^{<k>} = p_{loss\_BAT}^{<k>} + p_{loss\_SC}^{<k>} \qquad (8)$$

unless

$$p_{loss}^{<k>}(x^{<k>}) = p_{loss\_min}^{<k>}$$

$$x_m^{<k>} = \arg\min_x (p_{loss}^{<k>}), x \in [0, 1]$$

$$p_{loss\_min}^{<k>} = p_{loss}^{<k>}(x_m^{<k>}) \qquad (10)$$

$$W_{loss}^{<k>} = W_{loss}^{<k-1>} + T_S p_{loss\_min}^{<k>} \qquad (11)$$
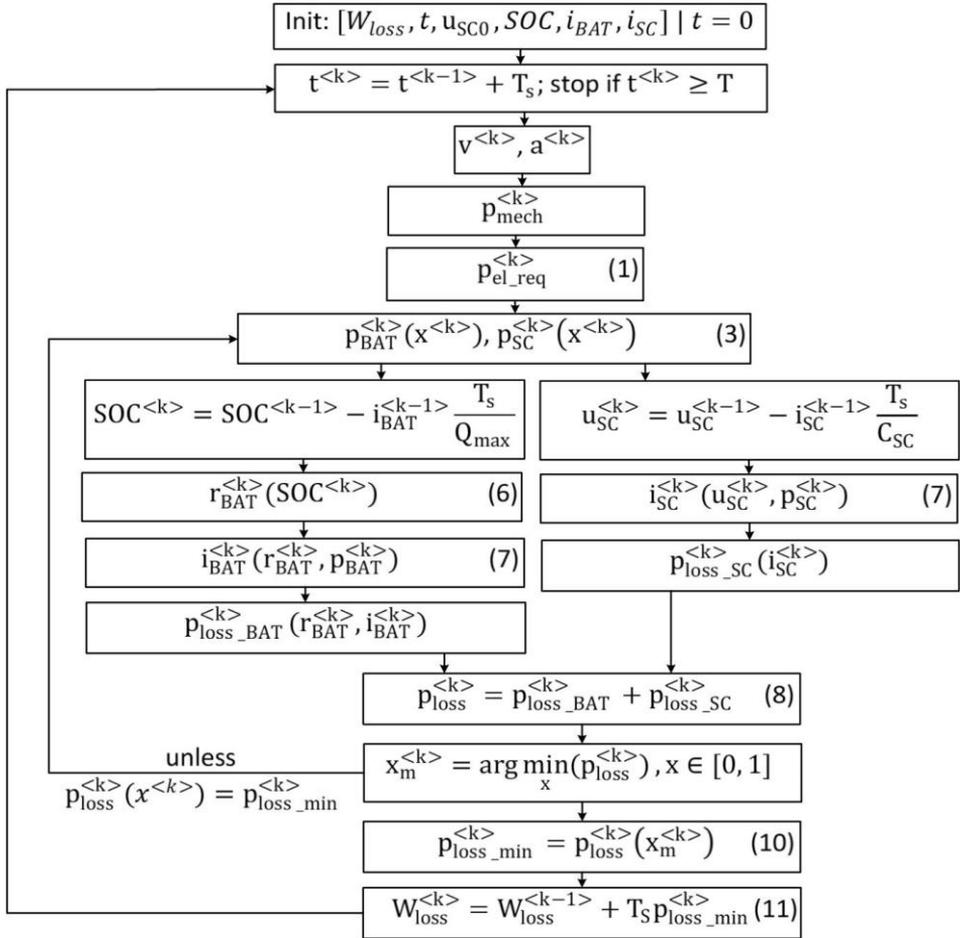
Figure 4: Flowchart in case of $J_p = p_{loss}(t)$

$$x = x_{init}$$

$$x = x + \Delta x \text{ (opt. alg.)}$$

Init: $[W_{loss}, u_{SC}, SOC, i_{BAT}, i_{SC}] \mid t = 0$

$$t^{<k>} = t^{<k-1>} + T_s$$

$$v^{<k>}, a^{<k>}$$

$$p_{mech}^{<k>}$$

$$p_{el\_req}^{<k>} \qquad (1)$$

$$p_{BAT}^{<k>}(x_i), p_{SC}^{<k>}(x_i) \ (3)$$

$$SOC^{<k>} = SOC^{<k-1>} - i_{BAT}^{<k-1>} \frac{T_s}{Q_{max}}$$

$$u_{SC}^{<k>} = u_{SC}^{<k-1>} - i_{SC}^{<k-1>} \frac{T_s}{C_{SC}}$$

$$r_{BAT}^{<k>}(SOC^{<k>}) \qquad (6)$$

$$i_{SC}^{<k>}(u_{SC}^{<k>}, p_{SC}^{<k>}(x_i)) \qquad (7)$$

$$i_{BAT}^{<k>}(r_{BAT}^{<k>}, p_{BAT}^{<k>}(x_i)) \qquad (7)$$

$$p_{loss\_SC}^{<k>}(i_{SC}^{<k>})$$

$$p_{loss\_BAT}^{<k>}(r_{BAT}^{<k>}, i_{BAT}^{<k>})$$

$$p_{loss}^{<k>} = p_{loss\_BAT}^{<k>} + p_{loss\_SC}^{<k>} \qquad (8)$$

$$W_{loss}^{<k>} = W_{loss}^{<k-1>} + T_S p_{loss}^{<k>} \qquad (9)$$

N $\qquad k \geq N = \dfrac{T}{T_s}$

Y

$$W_{loss}^{<N>}$$

$$x_m = \arg\min_x(W_{loss}^{<N>}), x_i \in [0, 1]$$

$$W_{loss\_min} = W_{loss}(x_m)$$
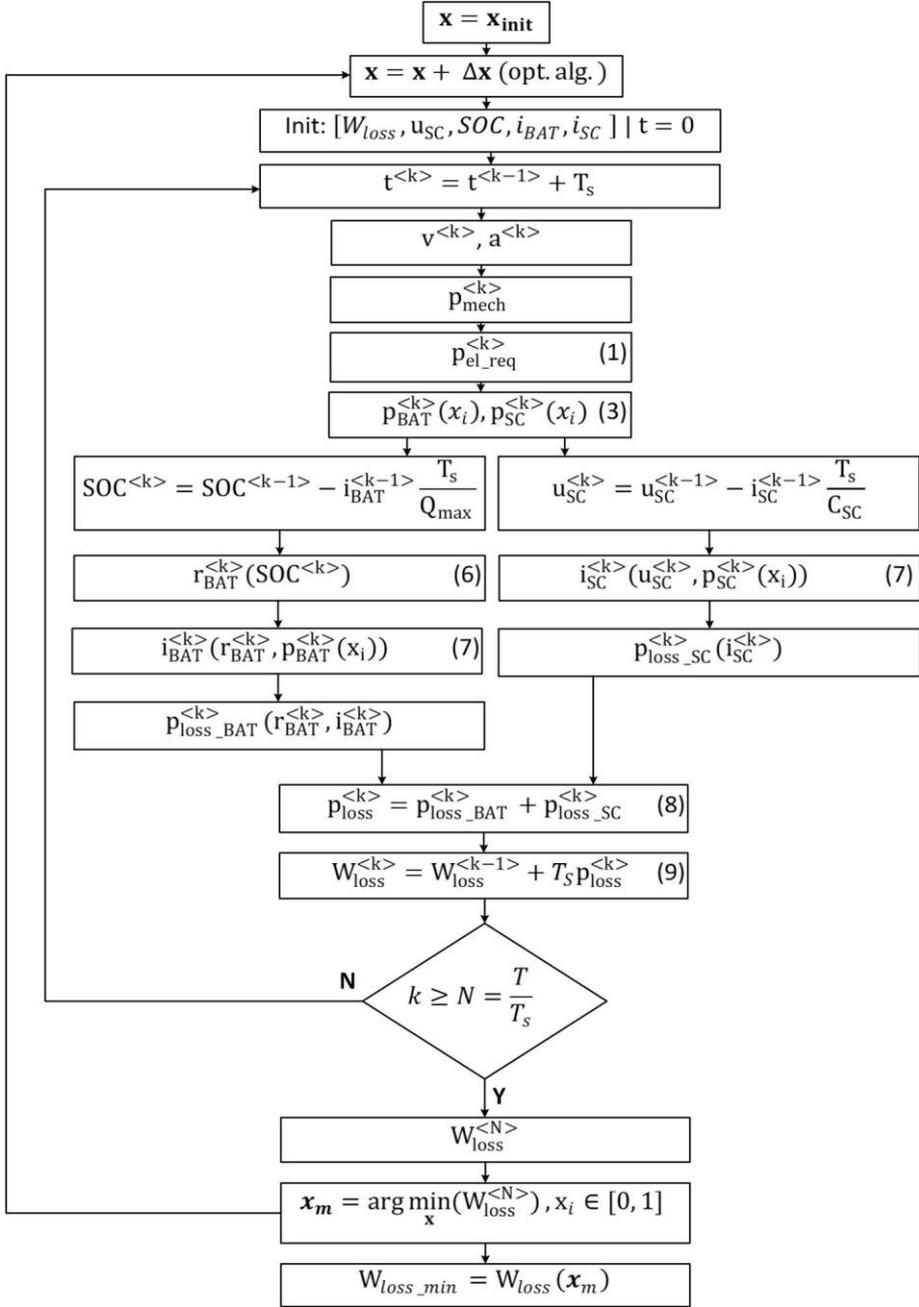
Figure 5: Flowchart in case of $J_W = W_{loss}(T)$

3. The particle swarm optimization algorithm

The particle swarm optimization method (PSO) has been chosen for the minimization of $J_W = W_{loss}(T)$ [12].

This intensely studied stochastic method [3], which simulates the social behaviour of animal swarms, is appied in this paper for simultaneous search over large regions of the solution space of the objective function. The search algorithm followed by an individual particle is based both on its own "experience" and on the results obtained by other particles from the swarm.

There exist several versions of the algorithm [5], [6], [15]. However, its principle can be formulated according to the relation (12), which defines the new position of the particle by using a "speed" vector (in fact a position increment) that contains components of randomized magnitudes. In case of an unconstrained optimization, these components are derived taking into account the initial speed of the particle (inertial term), its best previous result (cognitive term) and the best previous result of the swarm (social term) [3].

$$\begin{cases} \boldsymbol{v}_i^{k+1} = \omega \boldsymbol{v}_i^k + c_1 \boldsymbol{rand}_{i1}^k \odot \left(\boldsymbol{r}_{Bi}^k - \boldsymbol{r}_i^k \right) + c_2 \boldsymbol{rand}_{i2}^k \odot \left(\boldsymbol{r}_G^k - \boldsymbol{r}_i^k \right) \\ \boldsymbol{r}_i^{k+1} = \boldsymbol{r}_i^k + \boldsymbol{v}_i^{k+1} \end{cases}, (12)$$

where the notations stand for:

$\boldsymbol{r}_i^k$ – position vector of particle i in the k-th step of the search;

$\boldsymbol{v}_i^k$ – "speed"of particle i in the k-th step of the search;

$\boldsymbol{r}_{Bi}^k$ – individual best position vector of particle i until the k-th step of the search;

$\boldsymbol{r}_G^k$ – best position vector of any particle from the swarm until the k-th step of the search;

$\boldsymbol{rand}_{i1}^k$ and $\boldsymbol{rand}_{i2}^k$ are random vectors, with elements with continuous uniform distribution, in the range [0,1];

ω – inertia weight;

c_1 – cognitive learning factor;

c_2 – social learning factor;

$\odot$ – Hadamard product of vectors.

The algorithm in a two-dimensional space is illustrated in *Fig. 6*.
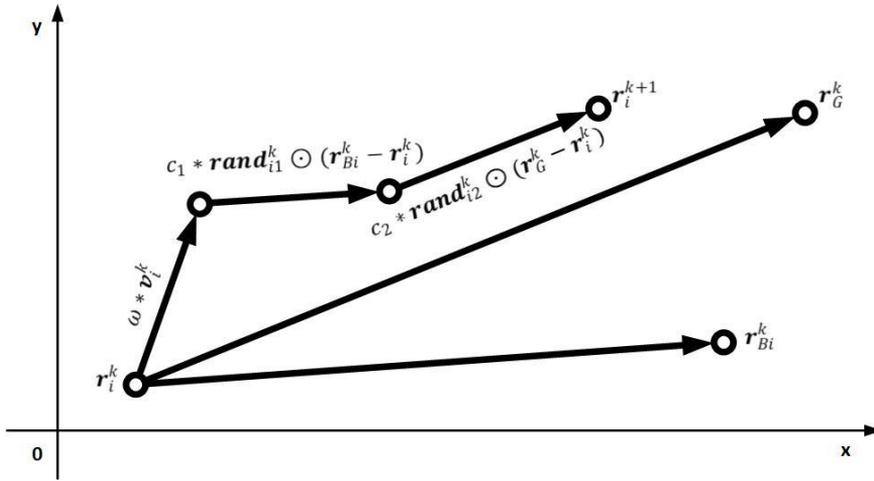
Figure 6: The unconstrained particle swarm optimization algorithm in a
two-dimensional space

   When the solution space is subject to constraints, the performance of swarm
optimization algorithm is largely influenced by the behaviour of the boundary,
especially when its position is close to the global optimum [2].

   For the application of inequality-type constraints that characterize the hybrid
energy storage system, in this study we analyzed the applicability of boundaries
with different behaviour, reported in the literature [2]. Namely, the "absorbing"
and "invisible" type boundaries have been verified for a standard test function,
and the "invisible" and a newly proposed "halving" behaviour have been
considered feasible in case of the HESS optimization.

   These are illustrated in *Fig. 7*. The principle of the "invisible" limitation
method is that if, according to (12), the particle should exit the solution space, the
new position is omitted from the calculation, and the position of that particle is
not refreshed. The "absorbing" method allows the motion of the particle in the
direction given by (12), but strictly limited to the surface of the boundary.

   The "halving" method preserves the direction of the speed vector (12), and it
iteratively halves its length until the new position of the particle fits the solution
space.

   The "absorbing" method (also the "reflecting" and "hybrid" mentioned in [2])
can be efficiently applied when the intersection point "A" in the bottom diagram
from *Fig. 7* can be determined easily. Unfortunately, this is not the case of the
HESS optimization.
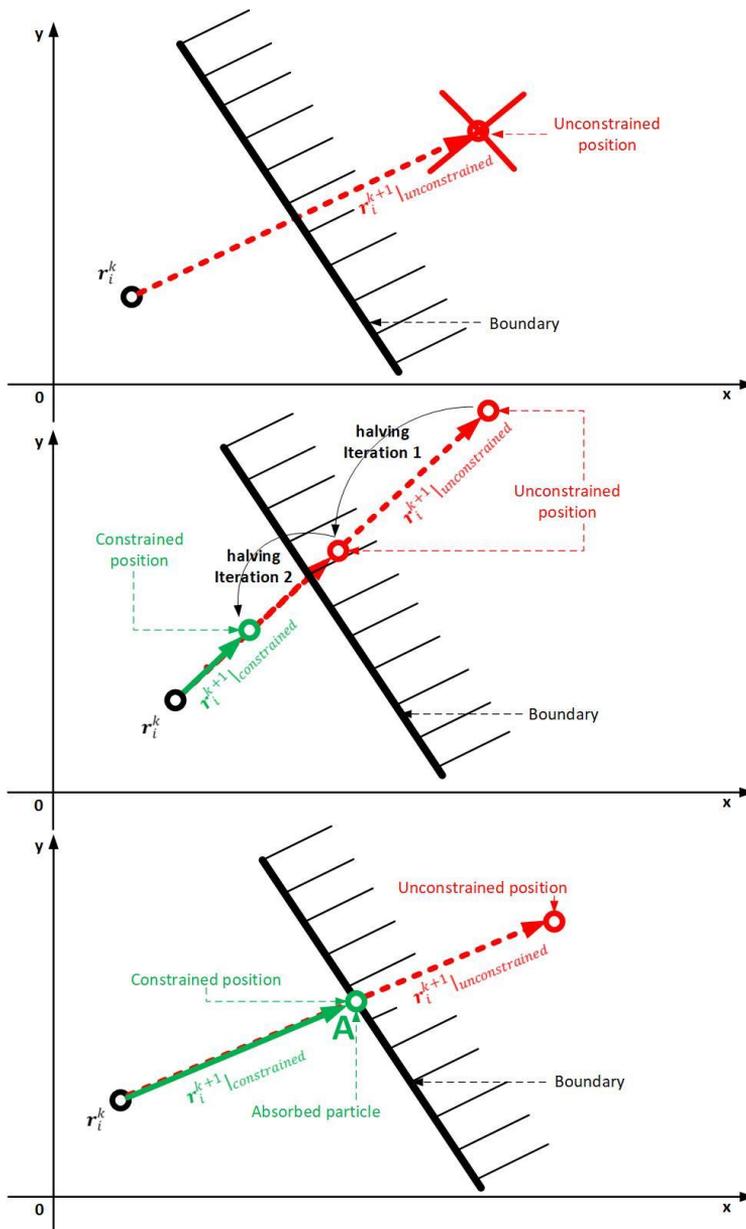
Figure 7: Illustration of the behaviour of the "invisible", "halving", and "absorbing", boundaries

A Matlab program has been developed to test the constrained swarm optimization algorithm applied to search the global minimum of the non-convex

Rastrigin function shown in *Fig. 8*, where the boundary is the green transparent plane.
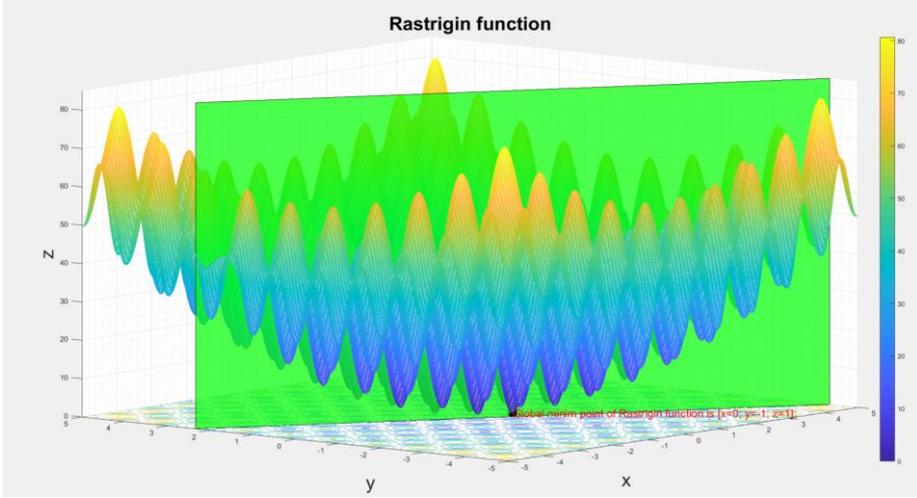
Figure 8: The Rastrigin function used for the performance test of the constrained optimization algorithm. The global minimum is set on the boundary

Table 1: The performance of the constrained optimization algorithm for different swarm sizes and boundary behaviours

Number of individuals in the swarm	Boundary behaviour	Number of searches out of 100, when only local minima were found	Average number of iterations for the cases when the global minimum was found
5	Absorption	90	146
5	Halving	94	114
5	Invisible	98	142
10	Absorption	32	73
10	Halving	56	130
10	Invisible	81	214
25	Absorption	3	54
25	Halving	5	131
25	Invisible	17	365
50	Absorption	1	54
50	Halving	3	130
50	Invisible	7	373
75	Absorption	1	52
75	Halving	1	139
75	Invisible	2	363

Table 1 summarizes the results of the constrained optimization algorithm applied to the problem from *Fig. 8*, obtained for different sizes of the swarm and different behaviours of the boundary.

The cognitive learning factor and the social learning factor were set to $c_1 = c_2 = 2$, while the inertia weight was $\omega = 0.073$.

Several 100 searches were performed for each swarm size and boundary type combination.

It can be concluded that a swarm of 5 particles is too small to cope with this task, as more than 90% of the randomly initialized searches fail to find the global minimum. It can be also observed that the increase of the number of particles above 25 does not determine the significant decrease of the number of iterations.

The statistics from *Fig. 9* demonstrate that the "absorbing" boundary is the most performant, with more than two times less iterations to the global minimum, than the "halving" one, and with 7 times less than the invisible" one.
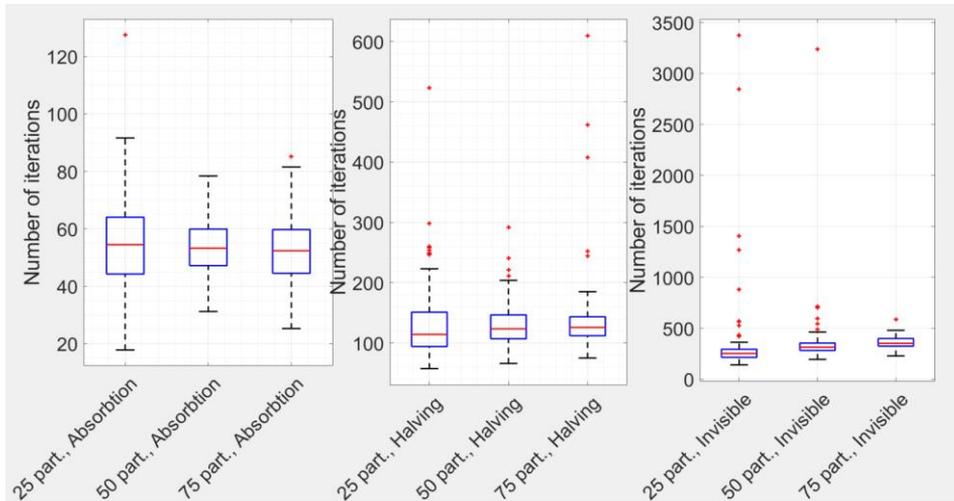
Figure 9: Boxplot statistics of the iteration number to the global minimum for different swarm sizes and boundary behaviours

4. Particle swarm minimization of the HESS energy losses

The solution of the energy loss minimization problem is represented by the time function of the power sharing ratio $x(t)$. Time discretization has been made according to *Fig. 5*, but it became evident that the dimensionality of the problem with the solution space of $x \in [0,1]^N$ is prohibitive even for a relatively short driving cycle.

Two approaches are presented in this section for the reduction of the dimensionality, applied both in case of unconstrained and constrained optimization.

The first approach consists of the division of the driving cycle to a small number of equal time intervals $P \ll \frac{T}{T_s}$, each with a constant value of the energy sharing factor x. Thus, the optimization task is reduced to a search in a P-dimensional solution space, i.e. the determination of a P-dimensional vector

$$\boldsymbol{x}_m = [x_1, x_2, \dots x_P] = \arg \min_{\boldsymbol{x}} \left(W_{loss}^{\langle N \rangle} \right), x_i \in [0,1]. \tag{13}$$

It is not totally surprising that such an unsupported subdivision of the driving cycle can yield results that are by far not optimal.

In the second version we propose a "dynamic" subdivision of the driving cycle. In this approach, the time instants that define the subdivision become dimensions of the solution space. Thus, the optimization task can be formulated as

$$\boldsymbol{x}_m^* = [\boldsymbol{x}_m, \boldsymbol{\tau}_m] = [x_1, x_2, \dots x_P, \tau_1, \tau_2, \dots \tau_{P-1}] = \arg \min_{\boldsymbol{x}^*} \left(W_{loss}^{\langle N \rangle} \right), x_i \in [0,1], \tau_i \in [0,1], \tag{14}$$

where $\tau_i = \frac{t_i}{T}$ are the subdivision time instants normalized to the driving cycle period.

The optimal solution is the vector $\boldsymbol{x}_m^*$, obtained by the extension of the optimal power sharing vector $\boldsymbol{x}_m$ with the normalized vector of the optimal subdivision time instants $\boldsymbol{\tau}_m$.

In this way the dimension of the solution space is almost doubled, but the complexity of the problem can still be handled for simple driving cycles.

We propose as an example of energy loss optimization the driving cycle from *Fig. 10*, which consists of a $T_a = 50\ s$, $a = 0.8\ \frac{m}{s^2}$ acceleration, and a $T_d = 50\ s$, $a = -0.8\ \frac{m}{s^2}$ deceleration stage of an $m = 1611\ kg$ vehicle. Thus, the maximum speed is $v_{max} = 144\ \frac{km}{h}$. The parameters of the energy storage devices are listed in *Table 2*. The capacity of the battery and the internal resistances have been distorted on purpose to obtain a large SOC variation, and to bring the HESS close to its operation limits in the short driving cycle from *Fig. 10*.
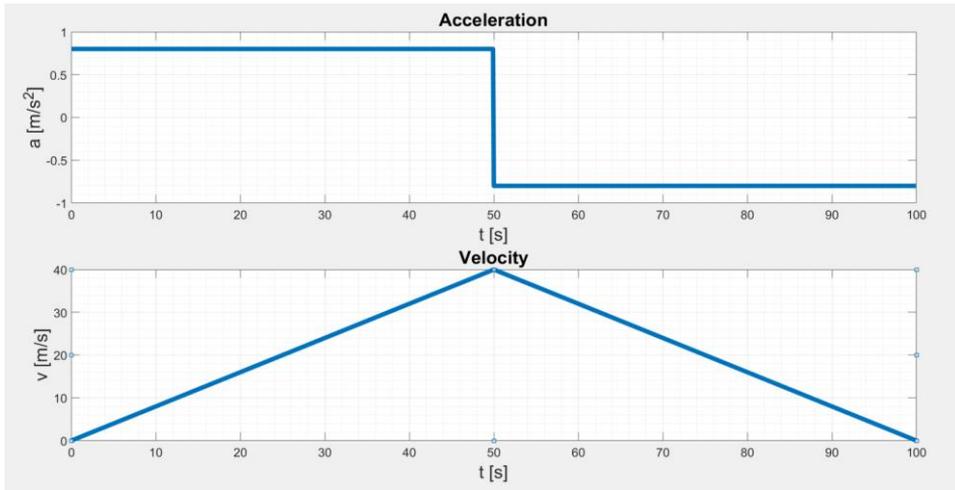
Figure 10: The simple driving cycle used to demonstrate the PSO of the total energy losses

Table 2: The parameters of the HESS, used for simulation.

Battery	Capacity	Q_{wh}	1000 Wh
	No load voltage	u_{BAT}	800 V
	Initial state of charge	SOC_{init}	100 %
	Internal resistance at SOC=100%	$r_{BAT}\|_{SOC=100\%}$	$300\ m\Omega$
	Internal resistance at SOC=50%	$r_{BAT}\|_{SOC=50\%}$	$650\ m\Omega$
Supercapacitor	Capacity	C_{SC}	10 F
	Initial voltage	U_{SC_init}	800 V
	Internal resistance	r_{SC}	$100\ m\Omega$

4.1 Unconstrained energy loss minimization

Fig. 11 shows the minimum values of the energy losses obtained by the application of the fixed and dynamic subdivision approaches, for $P = \{1, 2, 3, 4\}$. To be noticed, that the fixed division of the driving cycle into three equal subintervals yields higher losses than if it was divided only in two.

On the contrary, the dynamic subdivision yields a monotonically efficiency increasing with the number of subdivision intervals.
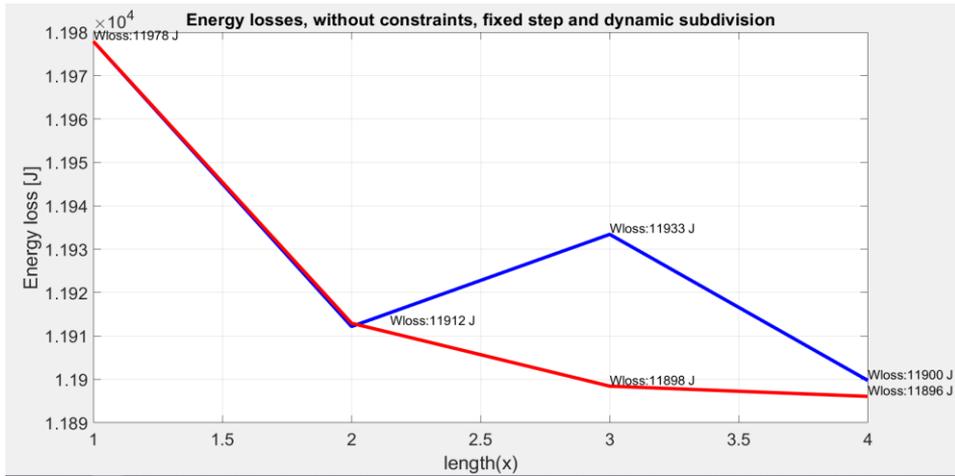
Figure 11: The results of the unconstrained PSO for fixed step (blue curve) and dynamic (red curve) subdivision of the driving cycle, versus the number of the subdivision intervals (the length of the power sharing vector)

Table 3 contains the optimal power sharing vectors and the optimal subdivision vectors for different numbers of subdivision intervals, resulted from the application of the fixed-step and dynamic subdivision methods.

Table 3: The results of the unconstrained energy loss optimization.

	Dynamic subdivision		Fixed step subdivision	
P	Optimal power sharing vector x_m	Optimal normalized subdivision vector τ_m	Optimal power sharing vector x_m	Optimal normalized subdivision vector τ_m
1	[0.6902]	-	[0.6902]	-
2	[0.6575, 0.7299]	[0.499]	[0.6604, 0.7290]	[0.5]
3	[0.5157, 0.6699 0.7317]	[0.159, 0.499]	[0.6391, 0.6936, 0.7416]	[0.3333, 0.6666]
4	[0.5389, 0.6499, 0.6663, 0.7235]	[0.131, 0.328, 0.503]	[0.6081, 0.6693, 0.7269, 0.7471]	[0.25, 0.5, 0.75]

4.2. Constrained energy loss minimization

The constrained particle swarm optimization of the HESS efficiency has been performed considering the limitation of the supercapacitor voltage:

$$U_{sc\_MIN} \leq u_{sc} \leq U_{sc\_MAX} , \qquad (15)$$

where $U_{sc\_MIN} = 0\ V, U_{sc\_MAX} = 800\ V$.

Fig. 11 and *Fig. 12* demonstrate that the dynamic subdivision method yields higher efficiency over the driving cycle than the fixed-step subdivision.
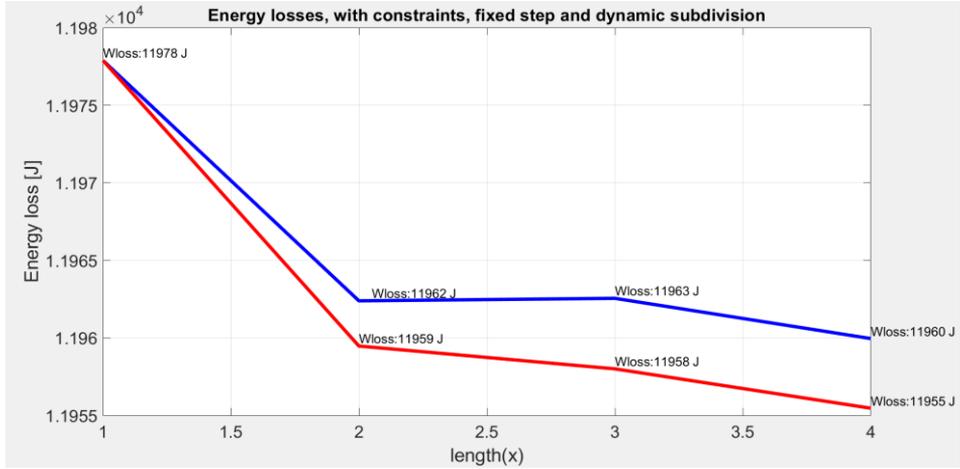
Figure 12: The results of the constrained PSO for fixed step (blue curve) and dynamic (red curve) subdivision of the driving cycle, versus the number of the subdivision intervals (the length of the power sharing vector)

The results for fixed-step and dynamic subdivision are shown in *Fig. 12*, while *Table 4* resumes the optimal power sharing vectors and the optimal subdivision vectors in case of the constrained PSO.

Fig. 13 illustrates the variation in time of different state variables of the HESS, when this is operated in the conditions that minimize the driving cycle energy losses. It can be observed that the constraints set for the supercapacitor voltage are satisfied.

Table 4: The results of the constrained energy loss optimization

P	Dynamic subdivision		Fixed step subdivision	
	Optimal power sharing vector x_m	Optimal normalized subdivision vector τ_m	Optimal power sharing vector x_m	Optimal normalized subdivision vector τ_m
1	[0.6902]	-	[0.6902]	-
2	[0.5344, 0.6917]	[0.1158]	[0.6868, 0.6955]	[0.5]
3	[0.5327, 0.6921 0.4877]	[0.1184, 0.9759]	[0.6781, 0.6925, 0.6976]	[0.3333, 0.6666]
4	[0.5067, 0.6579, 0.6941, 0.5535]	[0.0921, 0.2148, 0.9228]	[0.6654, 0.6909, 0.6967, 0.6825]	[0.25, 0.5, 0.75]
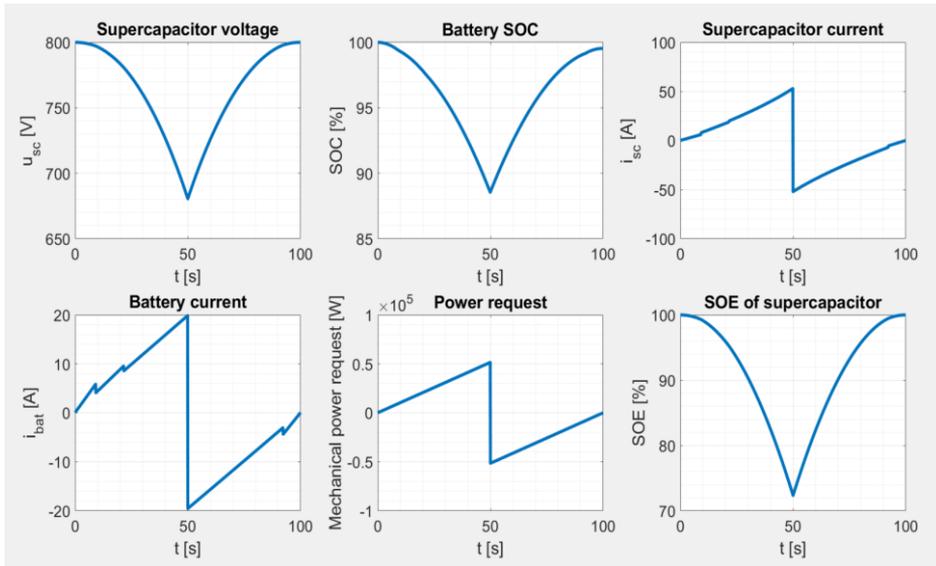
Figure 13: Time diagrams of the HESS state variables in case of operation under the condition $x^* = x_m^*$

4.3. Performance comparison of the "Invisible" and "Halving" boundaries

The influence of the boundary behaviour on the performance of the constrained optimization algorithm has been analyzed in case of the dynamic subdivision of the driving cycle to four time-intervals.
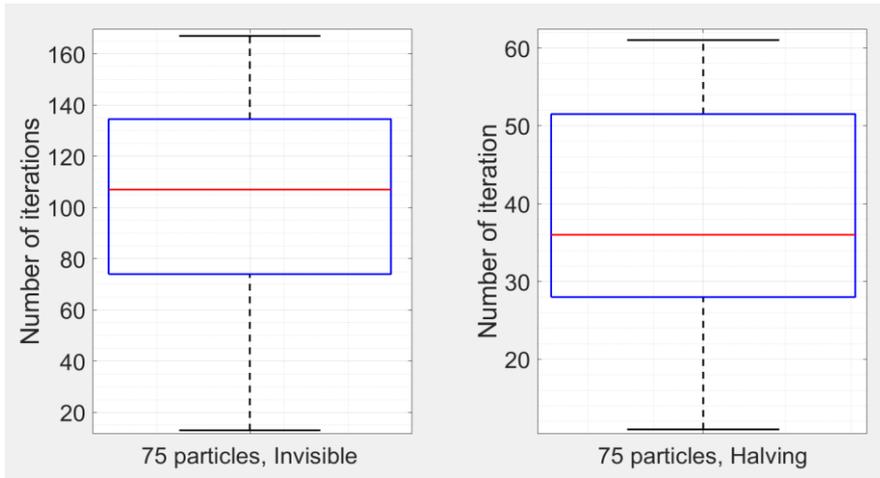
Figure 14: Boxplot statistics of the iteration number to the global minimum for a swarm of 75 particles in case of the "Invisible" and "Halving" boundaries

Thus, the dimension of the power sharing vector x is $P = 4$, and the dimension of the vector τ of the subdivision time instants is 3.

In *Fig. 14* it can be observed that the "Invisible" behaviour of the boundary results in almost three times more iterations to reach the global minimum, than the "Halving" behaviour. The statistics resulted from 20 search processes for each type of boundary, starting from different, randomly chosen initial positions.

5. Conclusions

In the study we aimed to minimize the losses of a hybrid energy storage system of an electric vehicle. This has been performed by means of an optimal sharing of the required electrical power between the energy storage devices. It has been shown that extending the dimension of the space of solutions with the number of the subdivision time instants of the driving cycle, the HESS can be efficiently optimized, while the dimensionality of the problem remains manageable.

The constrained optimization algorithm has been validated using the Rastrigin test function and three different behaviours of the boundary. Out of these, the "absorption" behaviour provided the best results, which is reasonable, especially in the cases when the optimum is close to the boundary.

For this test function, and a swarm of 75 particles, we found that the "Absorption" method is almost 7 times faster, and the "Halving" method is almost 3 times faster in finding the global minimum, than the "Invisible" method.

Unfortunately, the "Absorption" method can't be applied for the constrained optimization of the hybrid energy storage system, because in this case the

boundary itself is hard to be defined in the space of the power sharing factors and subdivision instants.

The local minima have been avoided by multiple initializations of the search algorithm, and combination of different stop conditions, including the clustering of the swarm, had to be applied in case of the constrained optimization.

References

[1]   Chen, Q., Jiang, J., Ruan, H., and Zhang, C., "Simply Designed and Universal Sliding Mode Observer for The SOC Estimation of Lithium-Ion Batteries", IET Power Electronics, vol. 10 Iss. 6, pp. 697−705, Jan. 2017.

[2]   Huang, T., and Mohan, A. S., "A Hybrid Boundary Condition for Robust Particle Swarm Optimization", *IEEE Antennas and Wireless Propagation Letters*, vol. 4, pp. 112−117, Jun. 2005.

[3]   Wang, D., Tan D., and Liu, L., "Particle Swarm Optimization Algorithm: An Overview", *Springer Link. Soft Computing*, vol. 22, pp. 387−408, Jan. 2017.

[4]   Kanglong, Y., Peiqing, L., and Hao, L., "Optimization of Hybrid Energy Storage System Control Strategy for Pure Electric Vehicle Based on Typical Driving Cycle", *Hindawi. Mathematical Problems in Engineering*, vol. 2020, pp. 1−12, Jun. 2020.

[5]   Chong, L., W., Wong, Y., W., and Rajkumar, R., K., "An Optimal Control Strategy for Standalone PV System with Battery Supercapacitor Hybrid Energy Storage System", *Elsevier. Journal of Power Sources*, vol. 331, pp. 553−565, Nov. 2016.

[6]   Ammari, C., Belatrache, D., Touhami, B., and Makhlouf, S., "Sizing, Optimization, Control and Energy Management of Hybrid Renewable Energy System — A Review", *KeAi. Energy and Built Environment*, vol. 2, pp. 23−53, Aug. 2021.

[7]   Romaus, C., Gathmann, K., and Böcker, J., "Optimal Energy Management for a Hybrid Energy Storage System for Electric Vehicles Based on Stochastic Dynamic Programming", *2010 IEEE. Vehicle Power and Propulsion Conference*, Sept. 2010, pp. 1−6.

[8]   Vinot, E., and Trigui, R., "Optimal Energy Management of HEVs with Hybrid Storage System", *HAL open science. Energy Conversion and Management*, vol. 76, pp. 437-452, Feb. 2019.

[9]   Capassoa, C., Lauria, D., and Veneria, O., "Optimal Control Strategy of Ultra-Capacitors in Hybrid Energy Storage System for Electric Vehicles", *Elsevier. Energy Procedia*, vol. 142, pp. 1914−1919, Aug. 2017.

[10]   Sadeq, T., Wai, C., K., Morris, E., Tarbosh, Q., A., and Aydoğdu, Ö., "Optimal Control Strategy to Maximize the Performance of Hybrid Energy Storage System for Electric Vehicle Considering Topography Information", *IEEE Access,* vol. 8, pp. 1−13, Nov. 2020.

[11]   Chang, W. Y., "The State of Charge Estimating Methods for Battery: A Review", *Hindawi. International Scholarly Research Notices*, vol. 2013, pp. 1-8, Jul. 2013.

[12]   Ye, K., and Li, P., "A New Adaptive PSO-PID Control Strategy of Hybrid Energy Storage System for Electric Vehicles", *Sage. Advances in Mechanical Engineering*, vol. 12, pp. 1−15, Aug. 2020.

[13]   Ferencz, J., Kelemen, A., and Imecs, M., "Control of an Electric Vehicle Hybrid Energy Storage System", *Sciendo, Papers on Technical Science*, vol. 14, pp. 77−88, Aug. 2021.

[14]   Paul, T., Mesbahi, T., and Durand, S., "Sizing of Lithium-Ion Battery/Supercapacitor Hybrid Energy Storage System for Forklift Vehicle", *Energies,* vol. 13, pp. 1−18, Sep. 2020.

[15]   Borgulya, I., "Optimalizálás evolúciós számításokkal", Typotex Kiadó Budapest, 2011.

Image Processing Methods for Gesture-Based Robot Control

Abdelouahab ZAATRI[1], Hamama ABOUD[2]

[1] Brothers Mentouri University Constantine1,
Department of Mechanical Engineering, Algeria,
e-mail:azaatri@yahoo.com
[2] Brothers Mentouri University Constantine1,
Department of Transportation Engineering, Algeria,
e-mail: hamamaaboud@yahoo.fr

Abstract: In this paper we discuss some image processing methods that can be used for motion recognition of human body parts such as hands or arms in order to interact with robots. This interaction is usually associated to gesture-based control. The considered image processing methods have been experienced for feature recognition in applications involving human robot interaction. They are namely: Sequential Similarity Detection Algorithm (SSDA), an appearance-based approach that uses image databases to model objects, and Kanade-Lucas-Tomasi (KLT) algorithm which is usually used for feature tracking. We illustrate the gesture-based interaction by using KLT algorithm. We discuss the adaptation of each of these methods to the context of gesture-based robot interaction and some of their related issues.

Keywords: Gesture-based control, human-robot interaction, feature recognition, movement tracking, robotics.

1. Introduction

In order to facilitate human-robot interaction and control, the recent design of various robots has fostered the emergence of several control modes. One can cite the following interaction modes: speech-based interaction, pointing on image-based interaction control, gesture-based control, myoelectric-based control, and even more recently brain-based control [1], [2].

As the many developed modes, gesture-based interaction has been studied, designed and implemented by many authors. A survey concerning gesture-based interaction is presented in [3], [4]. Technically, in most applications, gesture-

based interaction is mainly based on object recognition and tracking approaches involving CCD camera sensors. Several image processing methods have been used for image recognition such as feature-based methods, gradient-based methods, learning methods, etc. [5].

In principle, in gesture-based interaction, any part of the human body can be used for interacting with robots such as hands, arms, heads, eyes, lips, cheeks. Any tool moved by the human can also be used such as pencils, flags, sticks, etc. Applications of gesture-based robot interactions are numerous and can be adapted and extended according to many needs and contexts. Gestural expressions can be employed in regions where the speech is useless. This can happen in noisy places, in underwater areas, and in empty space where the medium cannot convey the sound waves as with astronauts [6]. It can be also used for learning by demonstration and for imitation as reproducing operations in medical care or tele-surgery [7]. This interaction mode is also interesting in some military activity to communicate and direct remote teams, autonomous and unmanned systems [8]. Moreover, it can be used in assistive robotics for supervising deaf people, for surveillance of disabled people, etc. As an example for such applications, we can cite one well known application that uses cheek movements to write and speak texts as in the case of the famous physicist Stephen Hawking [9], [11].

In addition, the gesture-based interaction can be designed with contact or without contact. Examples of gesture interaction with contact is used for teaching on blackboards, tables and other supports. In these situations, markers and colors can be employed to facilitate the tracking of elements of interest [10], [11]. However, with the advent of Covid19 pandemic, gesture-based control without contact is gaining a special importance by avoiding touch and contact with contaminated people and objects like door handles, machines, etc.

In this paper we discuss the human movement detection with applications oriented to gesture-based interactions. Some image processing methods that can serve as a core in this type of applications will be presented and discussed. We have largely applied them in object recognition and tested them with various robotics systems [11], [12], [13].

The first proposed method is a feature-based algorithm generally employed for stereo-vision to ensure correspondence between features in different images. This technique is called Sequential Similarity Detection Algorithm (SSDA) [14]. The second method belongs to recognition systems which are appearance-based. It uses a database of images to constitute models [15]. The last method uses the Kanade-Lucas-Tomasi (KLT) algorithm which belongs to gradient based trackers [16]. We will give briefly an illustration of gesture-based interaction experimented with KLT algorithm.

2. Description of some object recognition methods

This section provides a brief description of the three techniques of image processing we have experienced in object recognition and tracking of image features. We describe also the way they can be adapted and used for gesture-based interaction and control.

A. Sequential Similarity Detection Algorithm (SSDA)

SSDA is one of the basic algorithms which is typically used for 3D applications involving stereovision [14]. By stereovision, we mean that from two different images containing an object of interest, one can extract the 3D coordinates of this object in real world. In fact, the 3D information is hidden in the disparity between a same feature point observable in left and right images which are acquired from two different positions with a known displacement. Finding the corresponding points in the left and right images is the main problem in stereovision. Many techniques attempt to solve this difficult problem but with limited success depending on the context such as variation of luminosity, structure of the object, etc. [14].

Technically SSDA uses a similarity measure that is the correlation of light intensity between the neighborhoods of the right and left corresponding points situated in two images. First, a feature point in the left image is detected and a window is generated surrounding it. Its corresponding point in the right image is constrained to lie on a horizontal line of the image which is called epipolar line. Since many points of the right image can be candidates, therefore a window is generated sequentially around any candidate point. Then, the sum of light intensity difference over the windows of the left point and any candidate of the right image is computed and then squared. This sum-of-squared difference is defined as an evaluation function of the similarity between the left point and the candidate points of the right image. The best candidate that matches the left point corresponding to the minimal value of the evaluation function D, is the corresponding point. Mathematically, the algorithm determines *min(D)*[14], where:

$$D = \frac{1}{\sigma_L^2} \sum_i^n \sum_j^m \left\{ I_L(i,j) - I_R(i,j) \right\}^2 \tag{1}$$

with

$$\sigma_L^2 = \frac{1}{m.n} \sum_i^n \sum_j^m \left\{ I_L(i,j) - \mu_L \right\} \tag{2}$$

where $I_L(i,j)$ and $I_R(i,j)$ are respectively the light intensities of the pixel in left and right images situated at location (i, j), n and m are the width and length of the window generated around the location (i, j), μ_L and σ_L are respectively the mean and variance of the light intensity of the window containing the point in the left image.

We have implemented the SSDA which is a feature-based algorithm for object recognition and localization in many robotics applications [11], [12], [15]. To implement SSDA for movement detection or gesture-based interaction, we have adapted SSDA to match features from a temporal sequence of images. As we are concerned with relatively slow motion of the body part, only a very limited set of images during the movement can be used. The images are taken at a regular period T. Once two images separated by T are acquired, the SSDA is processed in order to determining the disparity in successive images. The disparity $disp(.)$ is defined as the difference between the positions of the feature point detected in right (u_{right}, v_{right}) and left (u_{left}, v_{left}) images in an image reference frame, such as:

$$disp(u) = u_{left} - u_{right} \tag{3}$$

$$disp(v) = v_{left} - v_{right} \tag{4}$$

To detect the occurrence of a movement of a body part, we do not need for this application to work in real word but it suffices to work in the image space. Therefore, we can infer the amplitude and the orientation of the motion, respectively.

This technique requires to specify in the initial static position one or more elements that constitute the model of the object to be tracked. To improve the efficiency of our system, we consider its sensitivity against detection of small movements which are not relevant and not intended as commands. These noisy movements can be caused by the imperfections of the practical used tools, by the changes in luminosity, by the normal shaking movements of the human body, etc. Errors can also be caused by the procedures that are inherent to the vision system itself. Examples are mismatches in determining corresponding points, mismatches in estimating orientations, etc.

To this end, an error detection technique has been defined based on a movement threshold value introduced to differentiate real movements from noisy ones. Then, correspondingly, threshold disparities are imposed to avoid the occurrence of noisy movements as follows:

$$disp(u) >= disp(u_{min}) \tag{5}$$

$$disp(v) >= disp(v_{min}) \tag{6}$$

where $u_{\min}$ and $v_{\min}$ are the thresholds values over which the movement can be considered as effective.

B. Object recognition system with an appearance based method

The existing approaches for image recognition in the literature are mainly of two types: model-based and appearance-basedmethods [15], [16], [17]. Model-based object recognition systems try to match a defined 3D model by its representation in a given image with 2D features such as lines, vertices and ellipses. Then, they try to extract from other images these features in order to recognise the objects. Appearance-based systems use the luminance information of an object. The idea is not to impose what has to be seen in the image (points, lines) but rather to use what is really seen in the image to characterise an object. In this respect, interest points are local features with high informational content. Examples are corners, T-junctions and locations where the texture varies significantly. Those interest points are on the other hand invariant with respect to geometric transformations such as rotations and translations. So that appearance-based systems learn objects by looking to them from different viewpoints and under different lighting conditions and use those images to model the object [15].

We briefly describe the recognition approach that belongs to appearance-based methods we have used and experienced. The corresponding software for object recognition which has been developed is described in [15]. This method adopts an invariant-based approach based on affinely invariant regions. Invariance is considered under affine geometric changes and linear photometric changes. This corresponds to the assumption that the scene is locally planar and not occluded and that no specular reflections occur. Precisely, the problem is as follows: given a point in two images of the same scene, taken from different viewpoints, find one or more regions around it, such that the same regions are found in both images independently, i.e. without knowledge about the other image. Correspondence between features is found by comparing affinely invariant regions. To reduce the complexity of the problem, restriction is made to finding affine invariant regions for corner points making use of the nearby edges [15].

In this implementation, the software is exploited as follows. First, a database of images containing the relevant objects (the models) is built off-line. The objects that will serve as models are placed into natural scenes. The operator moves his body part let's say his hand in a particular manner according to a given code which corresponds to a particular robot command. A set of images are captured from different views. The images are then processed in order to extract the features of the objects of interest that serve as models for the

automatic object recognition. To make this database more relevant, a region of enclosing the object of interest can be defined.

The process of feature extraction uses canny edge detector, corner detection and region finding. From the region finding invariant features concerning the model are extracted and are stored into the database. A model in the database may be represented in many scenes, under different viewpoints. The vision system takes a current image. This image is processed by the same procedures used to build the database models. The features of the objects are extracted from the current image. Then, the recognition process starts. It aims to match an object among those found in the current image to the one retrieved from the database.

C. Kanade–Lucas–Tomasi (KLT) feature tracker

The KLT algorithm is one of the most popular methods for feature tracking which was introduced by Lucas and Kanade [18] and later extended in the works of Tomasi and Kanade [19], [20]. As the SSDA, its main goal is to find corresponding features in different images. KLT makes use of spatial intensity information to direct the search for the position that yields the best match.

It uses an optical flow approach. The KLT algorithm proceeds in two steps. In the first step, it automatically detects a set of feature points in the initial image which have sufficient texture such as corners. These features are considered as best trackable in next images. In the second step, as with the SSDA, the correspondence is established by minimizing a function which is the sum of squared distances that measures the dissimilarity between a selected point and its potential corresponding one in the next image. By iterating this process sequentially between images, the features can be tracked. The final objective of the algorithm is to provide the coordinates of a feature in the sequence of images. Static features in images have a negligible displacement while moving features can be detected through their relative displacements.

KLT is a framework with the source code made available in the public domain, for both commercial and non-commercial use for the computer vision community. It is implemented in some platforms [21], [22]. The use of this software requires to specify some parameters among which the maximum number of tracked features needed for the application under consideration. This number is automatically detected in the first image if possible depending on the image structure and context. Nevertheless, in the next images, this number may decrease depending on the encountered situation since some features could be lost by occlusion, variations of luminosity, change in object orientation, change in distance of the object from the camera, etc.

Once the software is launched, the selected features are listed and colored in the initial images and in the next images the recognized features are colored as

well. This software is suitable for movement detection because the calculation of displacements of a feature in the sequence of frames enables consequently to detect the movement of the tracked object and its direction and orientation. One main advantage of the KLT is the fact that it does not need necessarily a specific model of an object to be tracked. The automatically detected features which are considered as trackable suffice to track the object of interest. Afterwards, in applications such as gestural interaction, it is primarily the movement that matters.

3. An example of application using KLT tracker

A. General structure of gesture-based robot control

The principle of gesture-based interaction systems is simple. It requires sensors that capture sequences of images during the movement of a human body part or of any tool moved by him/her. Then a software analyzes these sequences of images in order to track the movements of the elements of interest. Once a significant movement is detected and tracked; then, the final configuration of the gesture is identified. It is interpreted with respect to a code that links detected movements with corresponding robot commands. This configuration of the gesture is finally sent as a robot command corresponding to a particular action or task to be performed by the robot. *Fig. 1* summarizes the principle of gesture-based robot interaction showing the sequence of the designed and implemented general operations.
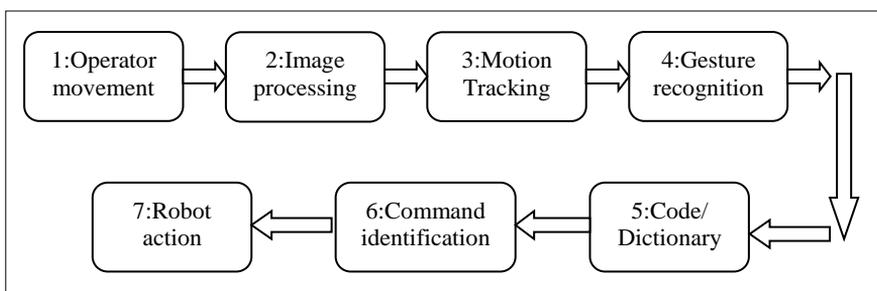
Figure 1: Sequence of involved general operations of gesture-based interaction

B. System description and experiments with KLT

Fig. 2 illustrates the general structure of one of our experimental system which was designed for gesture-based robot control and which was

implemented around KLT software [13]. On the right side of *Fig.2*, one can distinguish the human operator who is assumed to move a part of his body (hand, arm, leg, face, etc.) in front of a CCD camera. The camera captures a sequence of images which is analyzed by the KLT software to identify and track the operator's movements. On the right side of *Fig. 2*, one can distinguish a homemade serial robot manipulator that has to perform the corresponding commands according actually to human hand gestures. A virtual robot structurally shaped as the real robot in virtual environment was also developed by means of Java 3D platform. It enables to perform experimentations of the gestural-based robot control by simulations.

Many groups of experiments have been carried out with the serial robot manipulator shown in *Fig. 2*. The user was located in front of the camera up to 60 cm. The user performed a series of gestures (hand movements) in different directions. A maximum number of 30 feature points have been defined in the initial image. The image sequences are introduced as inputs to the KLT software which tracks the displacement of the identified features in successive images. The image coordinates of the tracked features (*u, v*) are identified. As we consider planar movements, these coordinates are converted via the well-known inverse perspective transformation into spatial coordinates (*x, y*). The direction of the movement is determined and consequently the robot moves its end effector to the corresponding directions. The command can be sent first for simulation to the virtual robot and then for execution to the real robot.
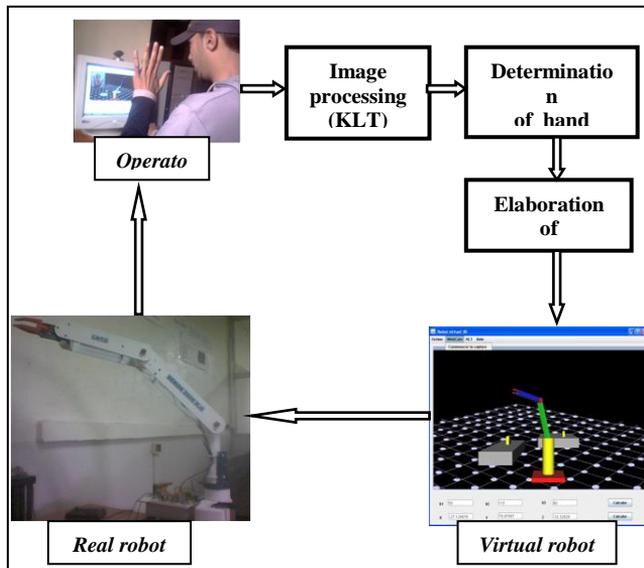
Figure 2: Experimental gesture-based system using KLT

The experiments have been carried out in two types of environments. In particular, seven groups of experiments have been performed in ordinary natural environments with different backgrounds. The environments contained a lot of objects in the background of the interesting object (user's hand) with possibly varying illumination. *Fig. 3.a* shows as an example of illustration two following images (at the left side of *Fig. 3.a*) captured by a CCD camera and analyzed by KLT. One can distinguish in the first image in red color the points to be tracked. In the following image, one can distinguish the points which have been recognized with many irrelevant elements which constitute a noise. Because of this, many points are lost in the second image and the success of the tracking decreases according to the complexity of the scene and of the environment. However, since we have only one object (hand) that is supposed to move, therefore very few corresponding points suffice to detect the movements. According to our experiments, about 60% were successful. Therefore, it appears necessary to control the environment to obtain more success and to make the gestural-based robot control effective.

A similar group of seven experiments has been performed in conditioned environments. In these experiments, the user's hand appears clearly without other objects in the background. The illumination is considered as stabilized. *Fig. 3.b* shows two successive images (at the right side of *Fig. 3.a*) captured during the execution of a hand movement. One can observe that the hand tracking is more successful compared to unconditioned and noisy environments. Experiments have clearly confirmed that conditioned environments highly improve the performance of the tracking process of the KLT algorithm. The success rate was about 90 %. The performance of this system can be improved also by using specific marks and colors to make the moving hand more easily detectable [11].
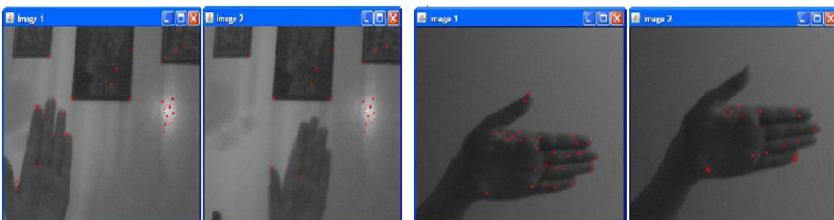
Figure 3: Experiments in different environments

a)-Natural environment (the two images on the left), b)-Conditioned environment (the two images on the right)

Considering the speed of the gesture-based interactions, some authors have tested the speed of KLT algorithm with different implementations on different platforms [23], [24]. Of course, it depends on the CPU, on the image size, on camera resolution, on the defined maximal number of features to be tracked, etc. In our case, we have used Java media framework to develop a program that extracts frames from the video sequence. The frames are of JPEG format and are then converted to PGM format. The camera resolution is 352*288 pixels. The process has been tested in real time to the KLT tracker. Since our applications require only relatively low speed motions, the execution of the image-processing based on KLT algorithm tested on ordinary computers in conditioned environments was about tenth of a second. Thus, this timing was considered as satisfying. However, it is possible to speed up real time applications that require faster execution time by specific equipment and optimization procedures [23], [24].

4. Discussion

In our analysis of this human-robot interaction mode, we consider two cases. The first case consists of focusing on tracking the movement. What counts here is not only the object to recognize as in static object recognition processes but it is the movement of objects itself. The second one consists of considering the configuration change of the object of interest once it stabilizes after the movement from the initial configuration to the final one. In fact, in this last case, the relevant information is contained in the final configuration that needs to be converted into a robot command according to a pre-established code.

All the described softwares can be adapted to track continuously in the video sequence the movement of the objects of interest. However, the problem of continuously tracking which is time consuming can be avoided in most applications since it suffices to detect the motion and to recognize the final configuration in order to identify the corresponding command to be issued. Again, according to our analysis, we inferred that each described image recognition method may be more efficient depending on the conditions of the application in its context. Here are some suggested cases to show how one can select the relevant method to apply with respect to the given contexts.

If we can define an easily recognizable structure that serves as a model of the human body part to be tracked (arm, hand, head, leg, etc.) and if the environment is not too complex; then the SSDA can be the best suitable software to apply.

If we have a limited set of movements that constitute a code and if the environment can be complex with possible variable illumination; then the appearance based method can be the best suitable software to apply. This

method can recognize the appropriate code after the end of the movement by simply comparing the actual obtained image with images contained in the database.

If the application does not require a specific model and the relevant information is contained in the movement itself and if the environment is not simple structured; therefore, KLT can be the best suitable software to apply as it does not require necessarily an initial specific model to detect. It tracks the movement of the human body part and when it stabilizes, it can identify the configuration to decode.

Of course, all these tracking vision methods are subject to some limitations, which are mainly due to the sensitivity of the vision systems, to lighting conditions, to the environment of work and to the application at hand. To improve the performance of the vision and tracking systems, our experiments have shown that it is recommended to condition the environment of work and to structure the tracked object. Moreover, other experiments have shown that using colors and markers may also improve the system performance by facilitating and guiding the tracking and recognition processes.

This gesture-based interaction mode was tested on various kinds of robots such as robot manipulators, mobile robots and cable-based robots as well [11]. Experimental tests showed that the gesture control mode can be effectively and worthily used for interaction with robot and telerobotics systems. New applications can be implemented where human are using gestures to interact with each other such as the take-off and landing of unmanned systems, aircraft fighters in military airport and marine vessels.

5. Conclusion

This article presents some methods of image processing which have been experimented for object recognition and 3D identification of locations. They are namely: Sequential Similarity Detection Algorithm (SSDA, Kanade-Lucas-Tomasi (KLT) algorithm, and an appearance-based approach that uses image databases to model objects. It discusses their applicability for designing and implementing gesture-based interaction with robots. An application using KLT tracker is presented.

Our analysis and experiments have shown that each of the presented methods can be adapted for a particular gestural system according to the body part used, the robotics application and the conditions of experiments. Of course, all these methods suffer from specific limitations, which are mainly due to the sensitivity of the vision system and of the environments. The conditioning of

the environment is an important factor that remarkably improves the performance of the vision system.

The experiments performed with various types of robots have confirmed the possibility of using this gesture detection mode as an interesting one for many applications in robotics and telerobotics domains. It can be also worthily included into multimodal human-robot interaction systems.

References

[1] Zaatri, A., and Ousalah, M., "Integration and design of multi-modal interfaces for supervisory control systems", *Journal of information fusion*, vol.4, pp. 135–150, 2003.

[2] Abiri, A., Heise, G., Zhao,X., Jiang,Y., and Abiri, F., "Brain Computer Interface for Gesture Control of a Social Robot: an Offline Study", in *Proc. 25th Iranian Conference on Electrical Engineering (ICEE), Tehran, Iran*, May 2-4, 2017, pp. 113–117.

[3] Mitra, S., and Acharya,T., "Gesture Recognition: A Survey", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp-311-324, May 2007.

[4] Sigalas, M, Baltzakis, H., and Trahanias, P., "Gesture recognition based on arm tracking for human-robot interaction", in *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems*, October 2010, pp. 5424–5429.

[5] Mathe, E., Mylonas, P., Spyrou, E., and Mylonas, P., "Arm Gesture Recognition using a Convolutional Neural Network", in Proc. *13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, 2018, pp. 37–42.

[6] Luo, J., Liu,Y.,and Ju. Z., "An Interactive Astronaut-Robot System with Gesture Control", *Comput. Intell. Neurosci.* vol.2016 (1), pp. 1-11, 2016.

[7] Staub, C., Can, C., Knoll, A., Nitsch,V., Karl, I., and Färber, B., "Implementation and Evaluation of a gesture-based Input Method in Robotic Surgery", in Proc. *IEEE International Workshop on Haptic Audio Visual Environments and Games (HAVE)*, Qinghuandao, Hebei, China, October 2011, pp.1–7.

[8] Elliott, R, E., Hill,S. G., and Barnes, M., "Gesture-Based Controls for Robots: Overview and Implications for Use by Soldiers", Human Research and Engineering Directorate, ARL, US Army Research Laboratory, ARL-TR-7715, July 2016.

[9] Kewley-Port, D., and Nearey, T. M., "Speech synthesizer produced voices for disabled, including Stephen Hawking", *The Journal of the Acoustical Society of America*, vol. 148, pp. R1-R2, 2020.

[10] Galván-Ruiz, J., Travieso-González, C.M., Tejera-Fettmilch, A., Pinan-Roescher, A., Esteban-Hernández, L., and Domínguez-Quintana, L., "Perspective and Evolution of Gesture Recognition for Sign Language: A Review", *Sensors (Basel)*. vol. 20, no. 12, pp. 3571, June 2020.

[11] Zaatri, A., and Bouchemal, B., "Some interactive control modes for planar cable-driven robots", *World Journal of Engineering*, vol. 10, no. 5, pp. 485–490, 2013.

[12] Zaatri, A., Tuytelaars, T., Waarsing, R., Van Brussel, H., and Van Gool, L., "Design and Implementation of a Supervised Intelligent Function", in Proc. *International Conference on Artificial Intelligence, IC-AI00*, *Las Vegas, USA*, 2000, pp. 799–805.

[13] Aboud, H., and Zaatri, A., "Suivi de gestes pour la commande des robots"*, in Proc. Congres Algerien de Mécanique (CAM-2011), Guelma*, November 2011, pp. 14–17.

[14] Shirai, Y., "Three-Dimensional Computer Vision", Springer.1987.

[15] Tuytelaars, T., Zaatri, A., Van Gool, L., and Van Brussel, "Automatic Object Recognition as Part of an Integrated Supervisory Control System", in Proc.*ICRA, San Francisco, USA*; 2000, pp. 3707–3712.

[16] Tuytelaars, T., and Mikolajczyk, K., "Local Invariant Feature Detectors: A Survey, Foundations and Trends", *Computer Graphics and Vision.*, vol. 3, no. 3, pp. 177–280, 2007.

[17] Baker, S., and Mattews, I., "Lucas-Kanade 20 Years On: A Unifying Framework", *International Journal of Computer Vision.,* Kluwer Academic Publishers, vol. 56, no. 3, pp. 221–255, 2004.

[18] Besl, P. J., and Jain, R. C., "Three-dimensional object recognition", *ACM Comput. Surv.*, (CSUR)17, pp.75–145, 1985.

[19] Lucas, B., and Kanade, T., "An iterative image registration technique with an application to stereo vision", in *Proceedings of the International Joint Conference on Artificial Intelligence*, 1981, pp. 674–679.

[20] Tomasi, C., and Kanade, T., "Shape and Motion from Image Streams: a Factorization Method—Part 3 -Detection and Tracking of Point Features", Technical Report CMU-CS-91–132, April 1991.

[21] https://cecas.clemson.edu/~stb/klt/

[22] http://www.java2s.com/ref/jar/download-klttracker131jar-file.html

[23] Fassold, H., et al., "Realtime KLT Feature Point Tracking for High Definition Video", Proceedings, https://www.yumpu.com/en/document/view/42917243/realtime-klt-feature-point-tracking-for-high-definition-semedia, 2009.

[24] Barnes, B., et al., "Evaluation of Feature Detectors for KLT based Feature Tracking using the Ondroid U3", in *Proceedings of Australasian Conference on Robotics and Automation*, The University of Melbourne, Melbourne, Australia, 2014, pp. 1–9.

Comparison of TCP SIAD and TCP BBR Congestion Control in Simulated 5G Networks

Donát Scharnitzky[1], Zsolt Krämer[2], Sándor Molnár[3]

Budapest University of Technology and Economics,
Faculty of Electrical Engineering and Informatics,
Department of Telecommunications and Media Informatics, Budapest, Hungary
e-mail: [1] scharnitzky@tmit.bme.hu, [2] kramer@tmit.bme.hu, [3] molnar@tmit.bme.hu

Abstract: 5G cellular networks have introduced a completely novel air interface called New Radio (NR). This technology delivers numerous benefits compared to previous generations, including significantly higher peak data rates. However, due to the propagation properties of the frequencies used in NR, the volatility of the available downlink capacity also increases. In this paper, we study two TCP congestion control algorithms which are designed to be able to quickly utilize sudden increases in available capacity. We present an implementation of TCP SIAD in the ns-3 open source network simulator and compare its performance with TCP BBR using the mmWave module of the simulator.

Keywords: TCP BBR, TCP SIAD, mmWave, 5G, ns-3

1. Introduction

The commercial deployment of 5G mobile networks is still ongoing, however, the benefits provided by the new air interface are already well established. New Radio (NR) is able to provide lower delay, better energy efficiency, and higher data rates compared to previous generations. Meanwhile, transport layer protocols and mechanisms also continue to evolve. Congestion control is a classic problem in transport layer performance optimization, and Cardwell et al. proposed a new algorithm in 2017 called Bottleneck Bandwidth and RTT (BBR) [1], which promised significant improvements. BBR has been compared to the current default congestion control algorithm in the Linux kernel, CUBIC in a number of recent studies.

In this paper, we study the interplay between congestion control algorithms and 5G mmWave environments. We present the implementation of TCP SIAD

(Scalable Increase, Adaptive Decrease), an algorithm with similar design goals as BBR in the ns-3 open source network simulator. Then we present the first performance comparison between the two algorithms in a scenario involving transitions between line of sight (LOS) and non-line-of-sight (NLOS) states.

The paper is organized as follows. In Section 2 we discuss the related work (namely TCP BBR, TCP SIAD and transport layer performance in 5G mmWave networks), then in Section 3 we describe the implementation of TCP SIAD in ns-3. Section 4 presents the simulation environment and the performance results, and Section 5 concludes the paper.

2.   Related work

A. TCP BBR

BBR was introduced by Google in 2017 in [1]. Compared to loss-based and delay-based congestion control algorithms, it represents a new approach, which can be called model-based, where the algorithm tries to maintain an estimation of the bottleneck bandwidth and RTT by active probing. The probing of bandwidth and RTT happens in separate phases, as probing for bandwidth increases latency and probing for RTT drains the queue, thus decreasing throughput. The other two phases of BBR's operation are called startup - and drain phases. The startup behavior is similar to the slow start in CUBIC. The drain phase after the startup tries to drain the additional queue at the bottleneck before the probing for bandwidth can begin.

A formal analysis and measurement study has been presented in [2] on the performance of BBR, which confirmed the intended behavior, however, also identified some cases where excess packet loss and fairness problems may occur. The authors of [3] compared the performance of BBR and CUBIC in real LTE networks on the highway, finding that while the achieved throughput was similar, BBR operated at significantly lower latency.

B. TCP SIAD

TCP Scalable Increase Adaptive Decrease [4] is a congestion control algorithm proposed by Kühlewind with the aim of having low delay while maintaining high utilization in different network conditions. The rate of feedback is independant of the bandwidth and can be manually set via a control point. To achieve this, the increase rate of the congestion window at congestion events is calculated in a way that the next congestion event is expected to happen after the same time as the previous one. This time is measured in the number of RTTs, and this is the configurable control point. The *epoch* is the time between two congestion events (congestion events in the same RTT are

regarded as the same). The buffers in the network have an impact on the delay, since congestions are needed for high utilization and they require the buffers to fill. TCP SIAD tries to empty the buffer every RTT when the delay is too large, which it can achive with reducing the congestion window. After the reduction the increase rate is recalculated to keep high utilization and congestion to happen for feedback (after the expected time since the last congestion event). The algorithm adapts to network environment changes, it detect new bandwidth with a similar method as Slow Start.

ISPs in a lot of cases set the buffer sizes to high values to reach high utilization [4], which leads to high delays and standing queues in case of packet loss based congestion control algorithms because the fillment of buffers. TCP SIAD is packet-loss based, but it also takes into account the delay, it avoids standing queues and reaches high utilizaton with small buffer sizes as well, thus the ISPs can lower the buffer sizes. The configurable control point allows external entities to control how aggressively the algorithm behaves, which makes it possible to apply higher level flow control mechanisms.

TCP SIAD consist of two main parts [4]: Scalable Increase and Adaptive Decrease. Scalable Increase calculates *alpha,* the rate of increase, in every *epoch* such that the *epoch*s will have the same length, but it does not modify *alpha* in an *epoch*. Adaptive Decrease calculates *beta*, the rate of decrease, based on the estimation of the number of packets in queue. TCP SIAD contains three additional algorithms: Fast Increase (to allocate new bandwidth), Additional Decrease (to empty buffers) and Trend Calculation (to improve convergence).

C. Transport layer performance in mmWave environments

The potential and challenges that 5G mmWave networks bring to the transport layer have motivated studies on the interplay between transport layer mechanisms and the characteristics of NR. A list of possible challenges are identified by the authors of [5], including rate adaptation, link quality judgement, bufferbloat, and beam misalignment. A more detailed analysis in [6] studies two deployment scenarios specified by 3GPP: high speed train and dense urban. The authors consider numerous factors that influence transport layer performance and thus different congestion control algorithms, TCP segment sizes, RLC (Radio Link Control) buffer sizes and server locations are compared. Regarding the server location, it is argued that the volatility of 5G networks increase the benefits of a shorter control loop for TCP. This is analyzed in detail in [7], presenting a comprehensive performance evaluation of transparent performance enhancing proxies in 5G mmWave networks. Regarding the different end-to-end congestion control algorithms, [6] finds that BBR significantly outperforms CUBIC, NewReno and HighSpeed in terms of

goodput, especially in the case of smaller RLC buffers. This holds true in both the remote server and the edge deployment cases.

One particularly interesting case in a NR environment is the one involving transitions between LOS and NLOS states. The performance of different TCP variants have been studied in great detail under these conditions. [8] shows that many TCP congestion control algorithms struggle to recover after the NLOS state, especially if the RTT is higher than 5 ms. The more aggressive algorithms (e.g., Scalable) achieved significantly higher throughput compared to CUBIC and Reno in these cases. [9] also investigates LOS-NLOS transitions in an urban deployment, and finds that in most configurations, BBR is able to outperform other TCP algorithms in terms of throughput and latency, and it was the only variant that could benefit from small buffers. In [10] the authors argue that for a 28GHz mmWave deployment, a 7MB RLC buffer results in optimal transport layer performance. A detailed investigation showed the effects of blockage on the different TCP variants, where BBR achieved lower latency than the loss-based algorithms.

All the aforementioned studies used the mmWave module of the ns-3 simulator, and assumed a 28 GHz frequency. An emulation-based measurement study presented in [11] assumed a 60 GHz frequency. Short blockages and long blockages were both studied and BBR avoided the large latency spikes experienced by CUBIC in both cases.

3.   Implementing TCP SIAD in ns3

Network Simulator 3 (ns-3) is an open source discrete time network simulation tool [12]. We used an extended version of ns-3, which contains a module for mmWave (5G) that can be used to add mmWave EPC, User Equipment, buildings, etc., to scenarios [13]. We used version 2.0 of the mmWave module, with the BBR implementation (and Internet module) from ns-3.35. Our implementation of TCP SIAD and the complete simulation setup can be found in a public repository [14].

A. Class hierarchy

TCP SIAD is implemented in the Linux kernel [15], we used it to guide our own implementation in ns-3. To add a congestion control algorithm to ns-3, one has to subclass the TcpCongestionOps class. Similarly how most of the congestion control algorithms use TcpNewReno as base class (which is a specialization of TcpCongestionOps) [16] we subclass TcpNewReno as can be seen in *Fig 1*. We added the new TCP SIAD class to the TypeId system for ease of access from scenarios.

TcpCongestionOps	TcpNewReno	TcpSiad
+ GetName()	+ GetName()	+ GetName()
+ CwndEvent()	+ Fork()	+ CwndEvent()
+ Fork()	+ IncreaseWindow()	+ Fork()
+ GetSshThresh()	+ GetTypeId()	+ GetSsthresh()
+ IncreaseWindow()	+ GetSsThresh()	+ IncreaseWindow()
+ PktsAcked()		+ PktsAcked()
+ GetTypeId()		+ GetTypeId()
+ Init()		
+ CongestionStateSet()		
+ HasCongControl()		
+ CongControl()		
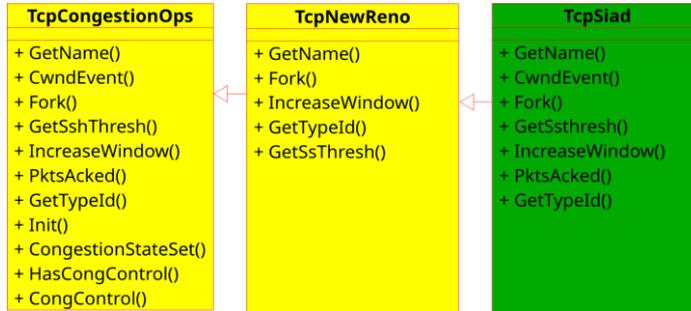
Figure 1: Class inheritance of congestion control, with the added TcpSiad class (green background)

B. Implementation details

We added a header and a source file to ns3 (these contain the SIAD implementation), other changes were not required. The algorithm is realized with four functions:

- GetSsThresh: calculates the new congestion window on a congestion event.
- IncreaseWindow: approximately called on ACKs, increases the congestion window.
- CwndEvent: resets the delay-related variables.
- PktsAcked: called after every ACK, determines the delay.

The variables are stored as class member variables, since the socket object that the functions get doesn't contain as much information as the Linux kernel's socket struct. *config_num_rtt* and *configNumMs* are registered by *TypeId* function, these variables provide the control loopback of TCP SIAD. The following two subchapters describe some implementation details.

C. Scalable Increase

The connection starts in Slow Start. If the congestion window is larger than *incthresh*, it needs to be grown via Fast Increase (Formula 1.) starting from the value 1 [4]:

$$alpha = alpha + \frac{alpha}{cwnd}, \; every \; ACK, alpha \leq \frac{cwnd}{2} \qquad (1)$$

If it is lesser than *ssthresh*, then via Slow Start. In these cases, there is no Additional Decrease, and this is indicated by setting *min_delay_seen* to true.

Additional Decrease can happen only after Adaptive Decrease or Additional Decrease (with approximately one RTT offset) in the linear increase phase.

If Additional Decrease is not needed, the congestion window is increased. *snd_cwnd_cnt* counts the segments, the counter is set to 0 when an increase happens, or when decrease happens (somewhere else).

alpha needs to be modified in the following four cases:

- We just entered linear increase phase after Slow Start (before the increase the congestion window is lesser than *ssthresh*, after that it is greater). *alpha* needs to be calculated according to Formula 2. [4]:

$$alpha = \frac{incthresh - ssthresh}{Num_{RTT}}, 1 < alpha < ssthresh \qquad (2)$$

- If we just left Slow Start, but there is no valid *incthresh* yet (*incthresh* < *sshthresh*) or we just left *incthresh* we switch to Fast Increase, thus *alpha* needs to be set to 1.
- We are in Fast Increase (congestion window is greater than *incthresh*). We add inc to *alpha* if it is less than half of the congestion window.
- We are in Slow Start (congestion window is lesser than *ssthresh*). *alpha* will always be the congestion window, with which we achieve to double the window.

D. Additional Decrease

Additional Decrease: *snd_cwnd_cnt* is set to 0 (see below). The congestion window is reduced according to Formula 3. [4]:

$$cwnd = \frac{RTT_{min}}{RTT_{curr}} ssthersh - 1 \qquad (3)$$

Then, if the congestion window is greater than the minimum congestion window, *alpha_new* (Formula 4.) and *red* (Formula 5.) is calculated [4]:

$$alpha_{new} = \frac{incthresh - cwnd}{Num_{RTT} - cnt_{dec} - 1} \qquad (4)$$

$$red = \frac{cwnd}{Num_{RTT} - cnt_{dec}} \qquad (5)$$

If *red* is greater than *alpha_new*, *alpha* is recalculated and the congestion window is reduced by *red*. If *red* is not greater that *alpha_new*, then *alpha* is set to *alpha_new* and is substracted from the congestion window. If the congestion window was the minimum congestion window, *alpha* has to be calculated

again. If at any point the congestion window becames minimal, *min_delay_seen* is set to true to disable the running of more Additional Decrease. If *alpha* is greater than the congestion window, then no more Additional Decreases happen (but *alpha* is not restricted, otherwise *incthresh* would not be reached in time).

4. Performance evaluation

A. *Simulation environment*

We created a scenario to test how TCP BBR and TCP SIAD behave in a volatile environment. It contains an EPC network which is connected to a remote host, this simulates a connection via the Internet. The topology has one eNodeB that is connected to one User Equipment (UE). The UE has a constant linear movement, after 5 seconds a building blocking its LOS to the eNodeB and then after another 5 seconds it goes back in LOS again.

We measured the received bytes and the RTT. *Table 1* shows the parameters that we changed from the default values. We compared TCP BBR and TCP SIAD congestion control algorithms in the same environment, where the environment was constant except for the RLC buffer size (and a random seed parameter).

Table 1: ns3 configuration parameters

Parameter	Value
Internet link RTT	25 ms
Internet link bandwidth	100 Gbps
Internet link MTU	1500 byte
TcpSocket segment size	10000 byte
TcpSocket min RTO	1000 ms
TcpSocket send buffer size	131072 * 50 byte
TcpSocket receive buffer size	131072 * 50 byte
AQM	Disabled
HARQ	Enabled
Center Frequency	28 GHz
Path loss model	BuildingsObstaclePropagationLossModel
Scheduler type	MmWaveFlexTtiMacScheduler
Simulation time	15 s
RLC mode	AM
RCL buffer size	[4, 7, 20, 40] MB

B. Performance results

Table 2 shows the Average Throughputs and RTTs for the different RLC buffer sizes and algorithms. At high buffers (7MB or greater), the congestion control algorithms behave virtually the same (they have a difference of less than 1%), thus having the same throughput and delay, which concludes that using more buffer than 7MB has no effect. *Fig 2* shows the utilization and fast adaptation of both algorithms in this case. Lowering the buffer however can negatively impact the throughput, especially in case of TCP BBR. TCP SIAD could still keep 90% of its utilization, and with this it achieved a 41% higher throughput than TCP BBR, while having the same delay.

At 4MB buffer the delay was 37% smaller compared to at 7MB buffers, which can justify using smaller buffers for delay sensitive applications.

Table 2: Average throughput and RTT for the different RLC buffer sizes

RLC buffer	Average Throughput (Mbps)		Average RTT (ms)	
	SIAD	BBR	SIAD	BBR
4 MB	1157.87	819.84	28.84	28.69
7 MB	1289.20	1295.10	46.03	45.81
20 MB	1290.57	1301.14	46.05	45.76
40 MB	1290.57	1301.14	46.05	45.76

When the UE moves behind the building, the throughput decreases and the delay increases significantly. After going in LOS again with the eNodeB, in case of higher buffer values, the throughput and delay get back to the previous values, as can be seen in *Fig 2*. In case of 4MB RLC buffer, after going back in LOS, TCP SIAD is able to utilize the bandwidth as well as before going in NLOS, however TCP BBR struggles to do so, as depicted in *Fig 3*.

It is worth highlighting that we have used an increased TCP MSS of 10KB based on previous studies [6] that showed how using larger segments can enable TCP to achieve higher throughput in mmWave environments. BBR however does not benefit from this due to its fundamentally different design.
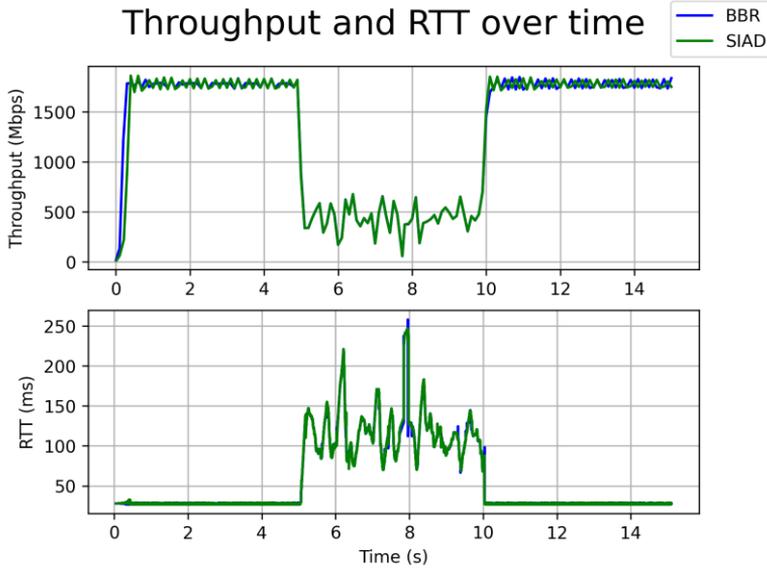
Figure 2: Throughput and delay of TCP BBR and TCP SIAD
at 7MB RLC buffer
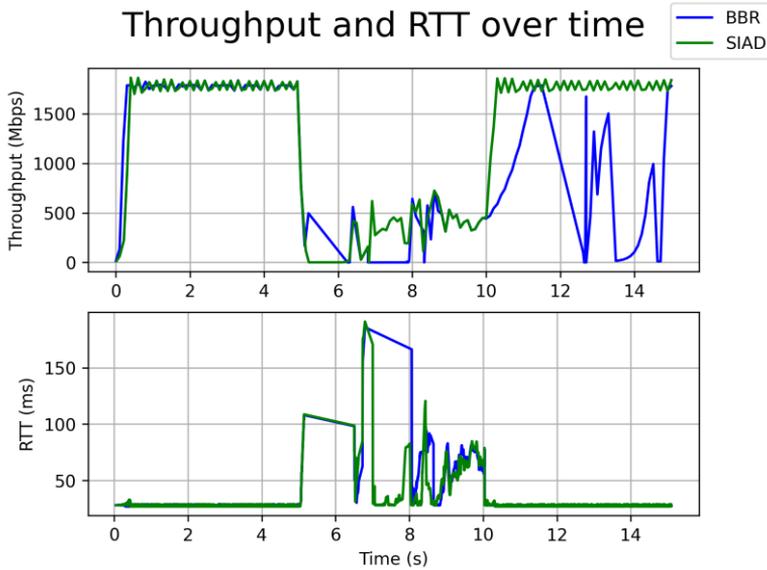
Figure 3: Throughput and delay of TCP BBR and TCP SIAD
at 4MB RLC buffer

5. Conclusion

The interplay between transport layer mechanisms and 5G mmWave network dynamics are complex and provide opportunities for optimizations. In this paper we have implemented TCP SIAD, an algorithm designed to quickly adapt to increased available bandwidth while keeping the delay low, and compared it to TCP BBR in a 5G mmWave environment using ns-3. We have shown that both algorithms perform well in terms of utilization and adaptation for RLC buffer sizes 7MB and higher, but at 4MB they start to struggle (especially TCP BBR). For low latency, however, 4MB buffer with TCP SIAD is a better choice, since it has only a small reduction in link utilization.

Our future work includes extending this analysis to different segment sizes and internet link RTTs.

References

[1] Cardwell, N., Cheng, Y., Stephen Gunn, C., Yeganeh, S. H., and Jacobson, V., "BBR: Congestion-Based Congestion Control", *Communications of the ACM* 60.2, 2017, pp. 58−66.

[2] Hock, M., Bless, R., and Zitterbart, M., "Experimental Evaluation of BBR Congestion Control", *IEEE 25th International Conference on Network Protocols (ICNP)*. IEEE, 2017.

[3] Feng, L., Chung, J. W., Jiang, X., and Claypool, M., "TCP CUBIC Versus BBR on the Highway", *International Conference on Passive and Active Network Measurement. Springer, Cham,* 2018.

[4] Kühlewind, M., "TCP SIAD: Congestion Control Supporting High Speed and Low Latency", arXiv preprint arXiv:1612.07947, 2016.

[5] Ren, Y., Yang, W., Zhou, X., Chen, H., Liu, B., "A Survey on TCP Over mmWave.", *Computer Communications,* 2021.

[6] Zhang, M., Polese, M., Mezzavilla, M., Zhu, J., Rangan, S., Panwar, S., Zorzi, M., "Will TCP Work in mmWave 5G Cellular Networks?", *IEEE Communications Magazine* 57.1, pp. 65−71, 2019.

[7] Hayes, D. A., Ros, D., and Alay, Ö., "On the Importance of TCP Splitting Proxies for Future 5G mmWave Communications", *IEEE 44th LCN Symposium on Emerging Topics in Networking (LCN Symposium, IEEE,* 2019.

[8] Pieska, M., and Kassler, A., "TCP Performance over 5G mmWave Links—Tradeoff Between Capacity and Latency", *IEEE 13th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), IEEE,* 2017.

[9] Poorzare, R., and Calveras Augé, A., "How Sufficient is TCP When Deployed in 5G mmWave Networks Over the Urban Deployment?", *IEEE Access 9,* 2021, 36342−36355.

[10] Mateo, P. J., Fiandrino, C., and Widmer, J., "Analysis of TCP Performance in 5G mm-wave Mobile Networks.", *ICC IEEE International Conference on Communications (ICC). IEEE,* 2019.

[11] Srivastava, A., Fund, F., and Panwar, S. S., "An Experimental Evaluation of Low Latency Congestion Control for mmwave Links", *IEEE INFOCOM Conference on Computer Communications Workshops (INFOCOM WKSHPS), IEEE,* 2020.

[12] https://www.nsnam.org/ [Online, last checked: 2021.11.18]

[13] Mezzavilla, M., Zhang, M., Polese, M., Ford, R., Dutta, S., Rangan, S., Zorzi, M., "End-to-End Simulation of 5G mmWave Networks," in *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2237–2263, 2018.

[14] https://github.com/dscharnitzky/mmwave-siad-bbr [Online, last checked: 2021.11.18]

[15] http://mirja.kuehlewind.net/src/tcp_siad.c [Online, last checked: 2021.11.18]

[16] https://www.nsnam.org/docs/release/3.35/doxygen/classns3_1_1_tcp_congestion_ops.html [Online, last checked: 2021.11.18]

Acta Universitatis Sapientiae

The scientific journal of Sapientia Hungarian University of Transylvania (Cluj-Napoca, Romania) publishes original papers and surveys in several areas of sciences written in English.
Information about each series can be found at
`http://www.acta.sapientia.ro`.

Main Editorial Board

László DÁVID Editor-in-Chief

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering

Executive Editor

Sapientia University

sciendo

Scientia Publishing House

ISSN 2065-5916
http://www.acta.sapientia.ro

Information for authors

Acta Universitatis Sapientiae, Electrical and Mechanical Engineering publishes only original papers and surveys in various fields of Electrical and Mechanical Engineering. All papers are peer-reviewed.

Papers published in current and previous volumes can be found in Portable Document Format (PDF) form at the address: `http://www.acta.sapientia.ro`.

The submitted papers must not be considered to be published by other journals. The corresponding author is responsible to obtain the permission for publication of co-authors and of the authorities of institutes, if needed. The Editorial Board is disclaiming any responsibility.

The paper must be submitted both in MSWord document and PDF format. The submitted PDF document is used as reference. The camera-ready journal is prepared in PDF format by the editors. In order to reduce subsequent changes of aspect to minimum, an accurate formatting is required. The paper should be prepared on A4 paper (210 × 297 mm) and it must contain an abstract of 200-250 words.

The language of the journal is English. The paper must be prepared in single-column format, not exceeding 12 pages including figures, tables and references.

For submission and the template file visit the following address:
`http://www.acta.sapientia.ro/acta-emeng/emeng-main.htm`.

One issue is offered to each author free of charge. No reprints are available.